Leibniz Universität Hannover

TIB Leibniz-Informationszentrum Technik und Naturwissenschaften Universitätsbibliothek

Gottfried Wilhelm Leibniz Universität Hannover
Fakultät für Elektrotechnik und Informatik

Development of an Analysis Process to Assess the Quality of Research Knowledge Graphs

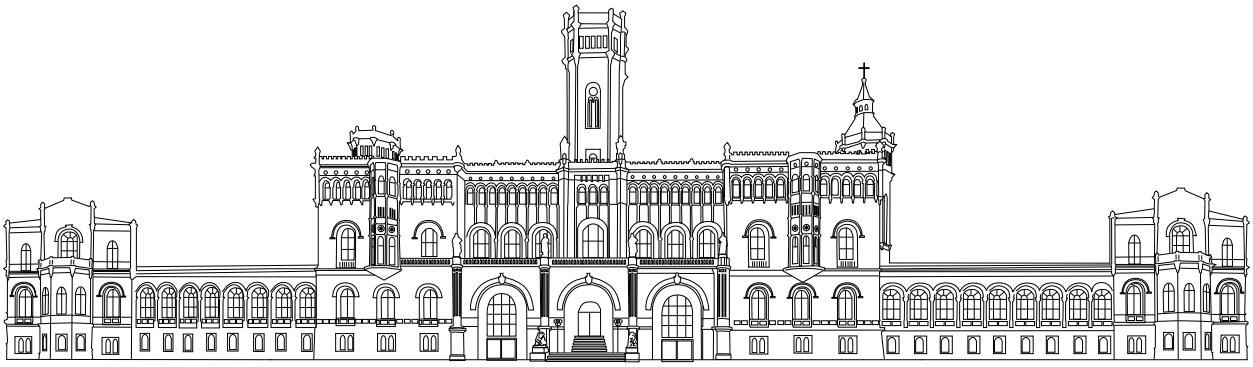
A thesis submitted in fulfillment of the requirements for the degree of
Master of Science in Computer Science

BY

Atiyeh Mohammadkhani
Matriculation number: 10027112
E-mail: Atiyeh.mohammadkhani@stud.uni-hannover.de

First evaluator: Prof. Dr. Sören Auer
Second evaluator: Dr. Oliver Karras
Supervisor: Dr. Oliver Karras, M.Sc. Hassan Hussein, M.Sc. Julia Evans

04. April 2023



Declaration of Authorship

I, Atiyeh Mohammadkhani, declare that this thesis titled, 'Development of an Analysis Process to Assess the Quality of Research Knowledge Graphs' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.

Atiyeh Mohammadkhani

04.04.2023

Abstract

This thesis proposes a novel approach for assessing the quality objectively of knowledge graphs, with a particular focus on the Open Research Knowledge Graph (ORKG). The ORKG is a community-driven open platform that aims to make research contributions more discoverable, accessible, and reusable. As a critical component of modern information systems, knowledge graphs enable effective data integration, discovery, and retrieval. However, assessing the quality of these graphs is challenging, given their complexity and heterogeneity.

The main problem addressed in this thesis is to develop an approach to assess the quality of knowledge graphs, with a particular emphasis on completeness and accuracy, in the context of the ORKG. The proposed approach is based on a set of quality measures that evaluate different aspects of completeness and accuracy, and it leverages the Knowledge Graph Maturity Model (KGMM) as a framework for assessing the maturity level of the ORKG.

The solution is evaluated empirically using a set of ORKG curation grants, and the observed results demonstrate that the proposed approach can effectively identify gaps in completeness and accuracy, and provide a comprehensive assessment of the quality of the ORKG. This assessment can help the ORKG community to prioritize curation efforts and improve the quality of the ORKG.

Overall, this thesis contributes to the field of knowledge graph quality assessment by proposing a comprehensive approach for assessing the quality of knowledge graphs, and demonstrating its effectiveness in the context of the ORKG. The proposed approach has the potential to be applied to other knowledge graphs, enabling better data integration, discovery, and retrieval in various domains.

Zusammenfassung

In dieser Arbeit wird ein neuartiger Ansatz zur objektiven Bewertung der Qualität von Wissensgraphen vorgeschlagen, wobei der Schwerpunkt auf dem Open Research Knowledge Graph (ORKG) liegt. Der ORKG ist eine von der Gemeinschaft betriebene offene Plattform, die darauf abzielt, Forschungsbeiträge besser auffindbar, zugänglich und wiederverwendbar zu machen. Wissensgraphen sind ein wichtiger Bestandteil moderner Informationssysteme und ermöglichen eine effektive Datenintegration, -suche und -abfrage. Die Bewertung der Qualität dieser Graphen ist jedoch angesichts ihrer Komplexität und Heterogenität eine Herausforderung.

Das Hauptproblem, das in dieser Arbeit behandelt wird, ist die Entwicklung eines Ansatzes zur Bewertung der Qualität von Wissensgraphen, mit besonderem Schwerpunkt auf Vollständigkeit und Genauigkeit, im Kontext des ORKG. Der vorgeschlagene Ansatz basiert auf einer Reihe von Qualitätsmaßstäben, die verschiedene Aspekte der Vollständigkeit und Genauigkeit bewerten, und er nutzt das Knowledge Graph Maturity Model (KGMM) als Rahmen für die Bewertung des Reifegrads des ORKG.

Die Lösung wird empirisch anhand einer Reihe von ORKG-Kuratoren evaluiert. Die Ergebnisse zeigen, dass der vorgeschlagene Ansatz effektiv Lücken in der Vollständigkeit und Genauigkeit identifizieren kann und eine umfassende Bewertung der Qualität des ORKG ermöglicht. Diese Bewertung kann der ORKG-Gemeinschaft helfen, Prioritäten bei der Kuratierung zu setzen und die Qualität des ORKG zu verbessern.

Insgesamt leistet diese Arbeit einen Beitrag zur Bewertung der Qualität von Wissensgraphen, indem sie einen umfassenden Ansatz zur Bewertung der Qualität von Wissensgraphen vorschlägt und dessen Wirksamkeit im Kontext des ORKG demonstriert. Der vorgeschlagene Ansatz hat das Potenzial, auf andere Wissensgraphen angewendet zu werden, um eine bessere Datenintegration, -suche und -abfrage in verschiedenen Bereichen zu ermöglichen.

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Dr. Oliver Karras, for his invaluable guidance, support, and encouragement throughout my thesis. His weekly meetings, constructive feedback, and insightful suggestions have greatly contributed to the completion of this work. I am also thankful to M.Sc. Hassan Hussein and M.Sc. Julia Evans for their helpful discussions and feedback. Their insights and recommendations have been instrumental in improving the quality and accuracy of my research.

Furthermore, I would like to extend my sincere appreciation to Prof. Dr. Sören Auer for allowing me to work with his research group in the context of the ORKG project. It has been an incredible opportunity to be a part of such a visionary project, and I am grateful for the knowledge and experience gained from this collaboration.

Last but not least, I would like to thank the ORKG Curation Team for their valuable feedback on my work.

I am also grateful to my family and friends for their unwavering support and encouragement throughout this journey. Their love and encouragement have been my source of strength, and I am forever indebted to them.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In this chapter, I define my aims and establish the fundamental purpose of my thesis. This chapter is organized as follows: Part 1.1 describes the fundamental motivator for the thesis project, whereas Section 1.2 details the precise aims I expect to achieve based on this motivation. Lastly, Section 1.3 discusses the thesis' general framework.

1.1 Motivation

Academic research has expanded rapidly in recent years, resulting in an increase in the number of research articles published. As a result, keeping up with the latest research is difficult for researchers. Finding relevant studies, comparing and replicating results, and having one's contributions recognized for their quality are all difficult tasks. Traditional scholarly communication methods, which rely on documents, are insufficient because they are incapable of organizing and presenting research knowledge in a machine-readable format. This method is inefficient because researchers must spend a significant amount of time searching for and interpreting study results from PDF files. It is a wasteful, time-consuming, and inefficient method of disseminating research findings.

The Open Research Knowledge Graph (ORKG)[1] represents a promising development in the field of scholarly communication. The ORKG aims to organize research articles using crowdsourcing, transforming scientific information into human- and machine-actionable knowledge. This platform enables whole new forms of machine aid, assisting researchers in locating important contributions to their domain and generating *comparisons* and *reviews*, which allows experts to compare different

[1] `https://orkg.org`

study publications so that similarities and differences between individual studies can be easily identified.This feature can help researchers make more informed decisions and gain more accurate insights, contributing to scientific knowledge.

The ORKG is a community-driven platform that aims to improve the quality of research communication. It provides researchers with the ability to explore knowledge in whole new ways, communicating discoveries across disciplines. By using the ORKG, researchers can share research findings in a machine-actionable format, making it easier for other researchers to discover and use the knowledge generated.

One of the ORKG's main advantages is that it allows academics to identify new correlations between study findings. By connecting relevant research articles, the ORKG can facilitate the identification of previously unknown ideas. The ORKG can also help researchers uncover relevant studies in their field of interest by improving the discoverability of research findings.

However, to realize the full potential of the ORKG and other research knowledge graphs, it is crucial to ensure the quality of the information presented on the platform. The quality of the research knowledge graph can have a considerable influence on the reliability and trustworthiness of the knowledge provided, as well as the usability and engagement of the research community.

The quality of knowledge graphs is a crucial issue in their effectiveness in various applications. While researchers have developed several quality assessment techniques and models to ensure the quality of knowledge graphs, the increasing complexity and volume of data have highlighted the need for a more comprehensive framework for assessing their maturity.

The Knowledge Graph Maturity Model (KGMM) [20] is a recent development that provides a structured and systematic approach to evaluating the quality and maturity of a knowledge graph. By assessing factors such as data completeness, accuracy, and consistency, the KGMM can help researchers identify areas for improvement and enhance the usability and engagement of research knowledge graphs, such as the Open Research Knowledge Graph (ORKG). This has the potential to contribute significantly to the advancement of scientific discovery and innovation. Therefore, the following research questions arises:

RQ1: How can the quality of the ORKG's comparisons be objectively evaluated using the Knowledge Graph Maturity Model (KGMM)?

RQ2: What insights can be gained by objectively evaluating the quality of the ORKG 's comparisons using the Knowledge Graph Maturity Model (KGMM)?

Ensuring the quality of research knowledge graphs is a complex issue that involves multiple factors. To improve research communication and facilitate the discovery of new insights, it is necessary to develop an analysis process that assesses individual

characteristics of knowledge graph quality. By identifying areas for improvement in these characteristics, we can enhance the overall quality of the knowledge graph, leading to increased usability and engagement by the research community

1.2 Goals

The goal of this thesis is to evaluate the quality of the graph-based comparisons in ORKG by considering a subset of characteristics from the Knowledge Graph Maturity Model (KGMM) [20].

The comparisons in ORKG are presented in the form of graph structures, and the focus is on evaluating the quality of these graphs. One way to measure the quality of knowledge graphs in ORKG is by analyzing the comparisons feature, which allows users to identify similarities and differences between individual contributions or research articles in the same field of study. By comparing contributions, researchers can gain a better understanding of the relationships between different pieces of information.

Furthermore, by analyzing comparisons, researchers can evaluate the accuracy, completeness, and consistency of the information presented in the ORKG. If many contributions in a comparison are found to be inconsistent or incomplete , this could indicate a problem with the overall quality of the knowledge graph in that domain. Conversely, if most contributions are consistent and complete, this would suggest a high-quality knowledge graph.

To reach this purpose, I consider the characteristics and measurements described in the KGMM, which divides measures into three categories: essential, important, and useful [20]. Essential measurements are crucial for creating a mature knowledge graph, whereas important measures aid in maturation and useful measures are not necessary but provide value to the knowledge graph.

In the Knowledge Graph Maturity Model (KGMM), levels refer to different stages of maturity in the development of a knowledge graph. There are five levels in the KGMM, each representing a higher degree of maturity and sophistication in the knowledge graph. At level 1, the knowledge graph should be made publicly available on the internet with an open license that satisfies certain quality criteria.

As ORKG has almost met the measures for Level 1, it can be said that it has achieved a basic level of maturity. However, to advance further, it's crucial to focus on the measures required for Level 2, which is called Completeness, and also Hussein [20] considers the presence of complete and up-to-date data as a crucial component of achieving maturity in knowledge graphs.

Level 2 includes 8 measures divided into two categories: essential and important. The essential ones are almost provided in ORKG, but four measures remain to be implemented. Specifically, I consider the measures of *instance completeness*, *population completeness*, *property completeness*, and *semantic accuracy*, which are categorized as important measures.

Based on the KGMM, assessing the quality of comparisons in ORKG through a (semi-)automated analysis process is the aim of this study. By evaluating the subset of measures defined in Level 2 of the KGMM, I can identify the strengths and weaknesses of the graphs and derive insights into the current quality. I conclude with recommendations on how the quality of the graphs can be improved to achieve a higher level of maturity in knowledge graphs. Overall, the goal of this thesis is to contribute to the improvement of the quality of graph-based comparisons in ORKG and to provide recommendations for future development of the platform. The outcome of this research will be a valuable resource for researchers and practitioners who are interested in graph-based comparisons and the assessment of the quality of knowledge graphs.

1.3 Structure

This thesis focuses on the quality assessment of knowledge graphs , which have become increasingly important in various domains. Chapter 2 provides an overview of knowledge graphs, their significance, and the state-of-the-art techniques and challenges associated with their quality assessment. This chapter also introduces the Knowledge Graph Maturity Model (KGMM) and evaluates its strengths and weaknesses. Chapter 3 covers related work on challenges, approaches, and techniques for knowledge graph quality assessment, as well as trends and future directions in this field.

Chapter 4 details the proposed approach for quality assessment of knowledge graphs, which includes completeness and accuracy as the main quality measures. This chapter explains how completeness is evaluated for property, instance, and population completeness, while semantic accuracy is assessed to ensure the correctness of the knowledge graph's relationships.

Chapter 5 presents the evaluation of quality measures, analysis of ORKG curation grants, and evaluation of completeness through different approaches. Finally, Chapter 6 discusses the strengths and limitations of the proposed approach and concludes with future work that can further enhance the quality assessment of knowledge graphs.

Chapter 2

Background

Chapter 2 of this thesis presents a comprehensive overview of the key concepts and frameworks in knowledge graph research. Section 2.1 explores the significance of knowledge graphs in modern data management systems, discussing their fundamental concepts, applications, and the role of ontology in their construction. This section emphasizes the importance of knowledge graphs in enabling efficient data integration and management, as well as enhanced data discovery and analysis.

In section 2.2, the field of Quality Assessment of Knowledge Graphs is examined, with a focus on state-of-the-art techniques and challenges facing researchers in this area. This section provides valuable context for understanding the evaluation methods employed in subsequent chapters.

Section 2.3 introduces the Knowledge Graph Maturity Model (KGMM) [20], a framework designed to evaluate and measure the quality of knowledge graphs. This framework is analyzed in section 2.3.1, providing a critical perspective on its strengths and weaknesses, as well as its practical applicability.

2.1 Knowledge Graphs

Knowledge graphs are a powerful form of structured knowledge representation that has gained significant attention in recent years. They provide a way to capture complex relationships and dependencies between entities in a specific domain or general knowledge domain, making it easier to analyze and reason over large volumes of data.

In this section, I provide an overview of knowledge graphs, including their importance and ontology.

2.1.1 Importance of Knowledge Graphs

Knowledge graphs are becoming increasingly important due to their ability to capture complex relationships and dependencies between entities in a domain. They have a broad range of applications across various fields, including natural language processing, information retrieval, data integration, and knowledge management[31]. For instance, in natural language processing, knowledge graphs can enhance the accuracy of named entity recognition, entity disambiguation, and relation extraction. In information retrieval, they can improve the relevance and diversity of search results by leveraging the semantic relationships between entities. In data integration, they can reconcile and integrate heterogeneous data sources by mapping them to a common knowledge representation[30].

According to Auer et al. [2], knowledge graphs provide a unified representation of data that can support a variety of use cases, including semantic search, question answering, and decision making. The authors emphasize the benefits of knowledge graphs in enabling more accurate and comprehensive analyses of data, as well as their potential to support automated reasoning and machine learning.

Ontology

As stated by Zhang et al. [45], the ontology of a knowledge graph refers to the set of concepts, relationships, and constraints that define the structure and semantics of the graph. In a knowledge graph, entities are represented as nodes, and the relationships between them are represented as edges. Nodes and edges can be enriched with additional attributes, such as labels, descriptions, and properties, which provide further context and semantics to the graph. The ontology of a knowledge graph can be formalized using standardized languages such as RDF, RDFS, and OWL, which enable representation and reasoning about the structure and semantics of the graph.

Euzenat and Shvaiko [8] offer a unified view of the ontology of knowledge graphs, emphasizing the importance of ontologies for knowledge representation. They argue that ontologies play a crucial role in making knowledge graphs interpretable, as they define the semantics of the graph and enable inference and reasoning. According to Euzenat and Shvaiko, the ontology of a knowledge graph should capture both the domain-specific knowledge and the general knowledge that is common to many domains.

In conclusion, knowledge graphs are a valuable tool for representing and organizing knowledge in a structured form, with many advantages over other knowledge representation techniques. The ontology of a knowledge graph defines the structure and semantics of the graph, and can be formalized using standardized languages such

as RDF, RDFS, and OWL. Resources such as Martin White's "Knowledge Graphs: An Introduction" [41] provide detailed explanations of the advantages and applications of knowledge graphs.

2.2 Quality Assessment of Knowledge Graphs: State-of-the-Art Techniques and Challenges

Quality assessment of knowledge graphs is an essential task to ensure their reliability, completeness, and consistency. Several state-of-the-art techniques have been proposed in recent years for evaluating the quality of knowledge graphs. As mentioned by [32], one of the main challenges in evaluating knowledge graphs is to identify relevant quality criteria, such as correctness, completeness, conciseness, and consistency, that reflect the intended use and domain of the knowledge graph. Different quality metrics and assessment methods have been proposed to measure these criteria, including precision, recall, F1-score, and various network analysis techniques[30].

Furthermore, as noted by [16], the quality assessment of knowledge graphs requires considering both the content and the structure of the graph. For instance, in addition to evaluating the correctness and completeness of the entities and relationships in the knowledge graph, it is crucial to assess the coherence and consistency of the graph structure, such as the connectivity, centrality, and clustering of the nodes and edges.

One of the major challenges in evaluating the quality of knowledge graphs is the lack of ground truth or reference data for comparison, as pointed out by [10]. This challenge has led to the development of several benchmarking frameworks and datasets for evaluating the quality of knowledge graphs, such as the Ontology Alignment Evaluation Initiative (OAEI) and the Knowledge Graph Analysis and Benchmarking (KG-BENCH) initiative.

Overall, the quality assessment of knowledge graphs is a complex and challenging task that requires considering various aspects of the graph's content and structure, as well as the intended use and domain. While several state-of-the-art techniques and benchmarking frameworks have been proposed, there is still much work to be done to address the open challenges and ensure the reliability and usefulness of knowledge graphs for various applications and domains.

2.3 Knowledge Graph Maturity Model (KGMM)

Knowledge graphs are becoming increasingly important in various domains and applications, enabling the representation of complex and heterogeneous data in a structured way. To ensure the quality of knowledge graphs, a structured approach for assessing their maturity is necessary. One such approach is the Knowledge Graph Maturity Model (KGMM) proposed by Hussein et al. in their paper titled "KGMM - A Maturity Model for Scholarly Knowledge Graphs based on Intertwined Human-Machine Collaboration" [20].

The KGMM consists of five maturity stages with 20 quality measures, which are prioritized in three categories in each level to support the applicability of the model. The model is inspired by the FAIR data principles [42], the Linked Open Data star scheme by Berners-Lee [3], and the Linked Data Quality Framework [44], but tailors and augments these frameworks specifically for scholarly knowledge graphs intertwining human-machine collaboration. The KGMM aims to ensure that knowledge graphs are reliable, accurate, and fit for purpose, enabling organizations to make more informed decisions and gain a competitive advantage in their respective fields.

The KGMM has been developed and is now available for use in large-scale scholarly knowledge graph curation efforts. It has shown promise in incrementally assessing and improving specific parts of the scholarly knowledge graph. Although it has only recently been released in production on the ORKG and its user statistics are unknown, the KGMM has the potential to enhance the quality and reliability of knowledge graphs. [20]

The KGMM provides clear guidelines for knowledge graph developers and curators to improve the maturity of their knowledge graphs, ensuring that the data is available for consumers in the most mature, complete, representable, stable, and linkable shape [19].

In conclusion, the KGMM provides a valuable tool for assessing the maturity of knowledge graphs in various domains and applications, enabling organizations to identify areas for improvement and allocate resources effectively. Its structured approach and prioritization of measures can help organizations ensure the quality of their knowledge graphs, ultimately leading to more informed decision-making and competitive advantage.

2.3.1 Evaluation of the Knowledge Graph Maturity Model: Strengths and Weaknesses

The Knowledge Graph Maturity Model (KGMM) has been proposed as a promising framework for evaluating the maturity of knowledge graphs across various domains and applications, although its adoption and impact have yet to be fully assessed due to its recent publication[19]. According to a study by Hussein et al. [20], the KGMM provides a structured approach for assessing the maturity of knowledge graphs and identifying areas for improvement, making it a valuable tool for organizations working with knowledge graphs.

One of the strengths of the KGMM is its comprehensive approach to evaluating the quality of knowledge graphs. As noted by Hussein et al. [20], the KGMM prioritizes measures by dividing them into three priorities: essential, important, and useful. This allows users to prioritize their efforts and allocate resources accordingly.

However, the KGMM has limitations that require attention. The KGMM tends to concentrate on the technical aspects of knowledge graph quality, overlooking the social and cultural factors that can affect knowledge graph development and usage. Jansen et al. [21] emphasized the importance of user engagement and knowledge sharing practices as crucial factors for the success of knowledge graphs, which are not adequately addressed by the KGMM.

Another limitation of the KGMM is that it does not account for the dynamic nature of knowledge graphs. As noted by Chen et al. [4], knowledge graphs are constantly evolving, and their quality may change over time. However, the KGMM does not provide a mechanism for monitoring and evaluating the quality of knowledge graphs over time, which can limit its usefulness for long-term knowledge graph management.

In conclusion, the KGMM is a valuable tool for evaluating the quality of knowledge graphs, but it has some limitations that need to be addressed. Future research should focus on developing more comprehensive frameworks that address the social and cultural factors that influence knowledge graph development and use, as well as mechanisms for monitoring and evaluating the quality of knowledge graphs over time.

Chapter 3

Related Work

The related work section of this thesis explores the challenges and various approaches to assessing the quality of knowledge graphs. First, the challenges related to knowledge graph quality assessment are discussed. Then, different techniques are reviewed, including metrics-based, heuristics-based, and benchmarks-based approaches. Finally, the section concludes by discussing trends and future directions in the field of knowledge graph quality assessment.

3.1 Challenges in Knowledge Graph Quality Assessment

The quality of a knowledge graph is crucial for its effective use and application. However, assessing the quality of a knowledge graph is a complex and challenging task. There are various factors that impact the quality of a knowledge graph, such as completeness, consistency, accuracy, and timeliness, among others. Measuring these factors and assessing the overall quality of a knowledge graph is not a straightforward process, and there is no one-size-fits-all solution to this problem.

One of the key challenges in quality assessment of knowledge graphs is the lack of a standardized and widely accepted framework for measuring quality. While there have been several proposals for quality assessment frameworks, there is still no consensus on what factors should be considered, how they should be measured, and how they should be weighted to compute an overall quality score for a knowledge graph. Moreover, the effectiveness of existing quality assessment methods and tools is still an open research question.

Another challenge in quality assessment of knowledge graphs is the scalability of

quality assessment methods. As knowledge graphs grow in size and complexity, the time and resources required to assess their quality increase significantly. Therefore, there is a need for scalable and efficient quality assessment methods and tools that can handle large-scale knowledge graphs.

To address these challenges and research questions, researchers have proposed various methods and frameworks for quality assessment of knowledge graphs. For example, the Knowledge Graph Maturity Model (KGMM) proposed by Hussein [20] provides a comprehensive framework for assessing the quality of scholarly knowledge graphs based on 20 measures and different categories for each measure, including completeness, consistency, accuracy, and timeliness. The framework also defines three priorities for quality assessment: essential, important, and useful.

Other approaches to quality assessment of knowledge graphs include the use of machine learning and natural language processing techniques to identify and correct errors in the knowledge graph [40], as well as the use of crowdsourcing and human-in-the-loop methods to improve the quality of the knowledge graph [23].

In summary, quality assessment of knowledge graphs is a challenging and important research area that requires further investigation. While there have been several proposals for quality assessment frameworks and methods, there is still a need for a standardized and widely accepted framework for measuring quality. Moreover, there is a need for scalable and efficient quality assessment methods and tools that can handle large-scale knowledge graphs.

3.2 Approaches and Techniques for Knowledge Graph Quality Assessment

In this section, we review and discuss some of the existing approaches and techniques for knowledge graph quality assessment. These include the use of metrics, heuristics, and benchmarks. We will compare and contrast these approaches, highlighting their strengths and limitations.

3.2.1 Metrics-based approaches

One common approach to knowledge graph quality assessment is the use of metrics. These metrics are typically quantitative measures that capture different aspects of a knowledge graph's quality, such as completeness, consistency, and accuracy. Some commonly used metrics include the number of triples, the number of unique entities and relations, and the frequency of specific relations or properties. Other metrics

capture more complex aspects of a knowledge graph, such as the coherence of its ontology or the degree of interlinkage between entities. Several studies have proposed different sets of metrics and benchmarks for knowledge graph quality assessment, such as the Luzzu framework [29] and the KG-Benchmarker [34]. While metrics-based approaches can provide valuable insights into a knowledge graph's quality, they have some limitations, such as the difficulty of defining appropriate metrics for specific domains or use cases.

One example of the difficulties of defining appropriate metrics for specific domains or use cases can be seen in the study by Singh et al. [34]. The authors noted that existing metrics may not be suitable for certain knowledge graph use cases, such as those involving natural language processing, and proposed a set of task-specific benchmarks to address this issue. This highlights the need for domain-specific metrics and benchmarks in knowledge graph quality assessment.

3.2.2 Heuristics-based approaches

Another approach to knowledge graph quality assessment is the use of heuristics, which are rules or guidelines that capture domain-specific knowledge or best practices. Heuristics-based approaches can complement metrics-based approaches, as they can provide more qualitative assessments of a knowledge graph's quality. For example, heuristics can be used to check the consistency of entity labels, the correctness of relationships between entities, or the adherence to a specific ontology or vocabulary[15].

Several studies have proposed different sets of heuristics for knowledge graph quality assessment, such as the 10 golden rules for knowledge graphs [15]. These rules include guidelines such as "use URIs as names for things" and "provide links to other datasets", which aim to improve the interoperability and reusability of knowledge graphs. Similarly, the quality dimensions for linked data proposed by Auer et al. [2] emphasize aspects such as "completeness", "accuracy", and "timeliness", which are essential for ensuring the quality and usefulness of linked data.

However, heuristics-based approaches also have some limitations, such as the subjectivity of the rules and the difficulty of defining comprehensive sets of heuristics for all possible use cases.

3.2.3 Benchmarks-based approaches

A third approach to knowledge graph quality assessment is the use of benchmarks, which are standardized datasets or tasks that can be used to evaluate the perfor-

mance of different knowledge graph systems or techniques. Benchmarks can provide a more realistic and comparable assessment of a knowledge graph's quality, as they can capture different aspects of its functionality, such as query answering, entity linking, or semantic reasoning. Some commonly used benchmarks for knowledge graph quality assessment include the Linking Open Data (LOD) dataset [2] and the Ontology Alignment Evaluation Initiative (OAEI) [9].

However, benchmarks-based approaches also have some limitations, such as the difficulty of defining comprehensive and representative datasets for all possible domains or use cases. While the OAEI benchmark covers a wide range of ontologies and alignment tasks, it may not be representative of all possible use cases, and its performance may vary depending on the specific domain or task. Similarly, the LOD dataset provides a rich source of interlinked data, but it may not be suitable for all types of knowledge graphs, such as those that require specialized or proprietary data sources.

3.3 Trends and Future Directions in Knowledge Graph Quality Assessment

I explore current trends and emerging research directions in knowledge graph quality assessment. These trends and directions have the potential to significantly improve the quality of knowledge graphs and make them more useful for various applications. One of the current trends in knowledge graph quality assessment is the integration of machine learning and natural language processing techniques. Machine learning techniques can be used to automatically learn quality patterns from existing knowledge graphs, and natural language processing techniques can be used to extract high-quality facts from unstructured text sources. For example, Wang [39] proposed a machine learning-based approach to identify incorrect relations in knowledge graphs, and Zhang [46] proposed a natural language processing-based approach to extract entity types from text sources.

Another trend in knowledge graph quality assessment is the use of crowdsourcing and human-in-the-loop approaches. Crowdsourcing can be used to collect high-quality annotations from a large number of users, and human-in-the-loop approaches can be used to incorporate human expertise and feedback in the quality assessment process. For example, Sun [36] proposed a crowdsourcing-based approach to verify the correctness of knowledge graph entities and relations, and Huma [17] proposed a human-in-the-loop approach to assess the completeness and consistency of knowledge graphs.

There is a rising interest in developing metrics and benchmarks that are specific to particular domains to ensure that quality assessment results are more precise and relevant. Domain-specific benchmarks can capture the unique features and requirements of different domains and applications. For instance, Schmachtenberg [33] proposed a domain-specific metric to evaluate the quality of geographic knowledge graphs, while Ma [25] introduced a benchmark for assessing the quality of biomedical knowledge graphs. Such domain-specific approaches to quality assessment can be highly effective in providing more accurate and relevant results.

In general, these emerging trends and directions can greatly enhance the quality of knowledge graphs and increase their usability in various applications. Nonetheless, there are still numerous challenges and research questions that need to be addressed. For instance, it is essential to explore effective methods of combining different quality assessment techniques and approaches. Additionally, it is important to develop ways of evaluating the uncertainty and incompleteness of knowledge graphs, as well as safeguarding the privacy and security of sensitive data during the quality assessment process.

In conclusion, in this section, we have discussed some of the current trends and emerging research directions in knowledge graph quality assessment. These trends and directions have the potential to significantly improve the quality of knowledge graphs and make them more useful for various applications. However, there are still many challenges and open research questions in this area, and further research is needed to address these challenges and develop more effective and efficient quality assessment approaches and techniques.

Chapter 4

Approach

The aim of this study is to evaluate the quality of graph-based comparisons in the Open Research Knowledge Graph (ORKG) using a subset of measures from the Knowledge Graph Maturity Model (KGMM) [20]. The focus is on four measures from the KGMM, namely population completeness, instance completeness, property completeness, and semantic accuracy .

In this section an overview of these measures and their mathematical formulations is provided.

4.1 Completeness

Completeness is a crucial aspect of knowledge graphs, as it ensures that all the relevant information is included and available for use. In the context of the Knowledge Graph Maturity Model (KGMM) [20], completeness is defined as the degree to which a knowledge graph includes all the relevant instances, properties, and relationships that are necessary to support its intended use cases. This includes ensuring that the knowledge graph is comprehensive, up-to-date, and accurate.

Completeness is essential for knowledge graphs used in various applications such as information retrieval, natural language processing, and data integration. Incomplete knowledge graphs can lead to erroneous or incomplete results in these applications, which can hinder decision-making processes or even result in incorrect conclusions. Therefore, it is important to ensure that knowledge graphs are complete and that their completeness is regularly evaluated and improved.

Overall, completeness is a fundamental aspect of knowledge graphs, and its importance cannot be overstated. By evaluating the completeness of knowledge graphs, it is possible to ensure that they are comprehensive, accurate, and up-to-date, which

can lead to better decision-making, improved data integration, and enhanced user experiences.

In the following sections, I delve deeper into three important components of completeness in the KGMM: property completeness, instance completeness, and population completeness.

4.1.1 Property Completeness

Property completeness is a measure of the extent to which all expected properties of a class or instance in a knowledge graph are represented. In other words, it measures the degree to which a knowledge graph covers all the properties that are relevant to the entities it represents. The concept of property completeness is closely related to the notion of schema completeness, which is the degree to which the schema of a knowledge graph covers all the relevant properties and relationships for the entities it represents.

Several methods have been proposed to measure property completeness in knowledge graphs. One common approach is to define a set of expected properties for a given class or instance based on domain knowledge or existing standards, and then compare this set to the actual set of properties represented in the knowledge graph. For example, Daga [6] propose a measure called Property Coverage, which is defined as the ratio of expected properties for a given class or instance that are represented in the knowledge graph. This measure can be used to identify gaps in the knowledge graph and prioritize efforts to improve its completeness.

Another approach to measuring property completeness is to use machine learning techniques to automatically identify relevant properties based on the available data. For example, in their work on the Open Research Knowledge Graph (ORKG), Hassan et al [13] use a machine learning model to predict the relevance of properties for a given class or instance, and then use this information to prioritize the representation of those properties in the knowledge graph. This approach can be particularly useful in cases where the domain knowledge is incomplete or where there are too many properties to manually specify.

Property Completeness in the ORKG

In the KGMM [20], property completeness is defined as the ratio of incomplete properties to the total number of properties in a knowledge graph. However, the ORKG realization of property completeness involves the completion of properties with the help of reviewers' suggestions, and the ultimate decision of whether to add a certain suggested property or not rests with the comparison author. This

approach acknowledges the collaborative nature of ORKG and the potential benefits of involving multiple reviewers in the completion of properties.

In order to measure property completeness in the ORKG, I consider it important that for each contribution in the comparison, there should be at least one resource or literal reference provided for each property. This ensures that each property has been considered and its relevance evaluated for each contribution in the comparison, and that there are no missing values or overlooked properties.By ensuring these criteria are met, we can confidently assert that the property completeness is at 100%.

Property completeness is calculated as follows in my approach. For each property, contributions are assigned a value of 1 if the contribution has at least one resource or literal for the given property, or 0 if it does not. The sum of these values is found and divided by the total number of contributions in the comparison to obtain the property completeness of each property.

I calculate this property completeness value for each property in a comparison, and then take the average across all properties to obtain the average property completeness score for the comparison.

The following formulas represent the calculations for property completeness based on the aforementioned criteria:

The formula for calculating the property completeness of a single property is given in Equation 1 where $P(C)$ is a property P with associated contribution C. If the contribution has one or more element(s) for the given property, the value of $P(C)$ is 1, otherwise it is 0. To calculate the average property completeness score for a comparison, the mean of all properties is taken, as shown in Equation 2.

$$f(P(C)) = \begin{cases} 1, & \text{if there is at least one element in P(C)} \\ 0, & \text{otherwise} \end{cases}$$

$$\frac{\sum_C f(P(C))}{|C|} \tag{1}$$

$$\frac{\sum_P \frac{\sum_C f(P(C))}{|C|}}{|P|} \tag{2}$$

This metric calculates what portion of properties have values across all contributions in a given comparison.

Here is an example of an ORKG comparison containing three contributions and

two properties, which demonstrates how to compute the property completeness metric.

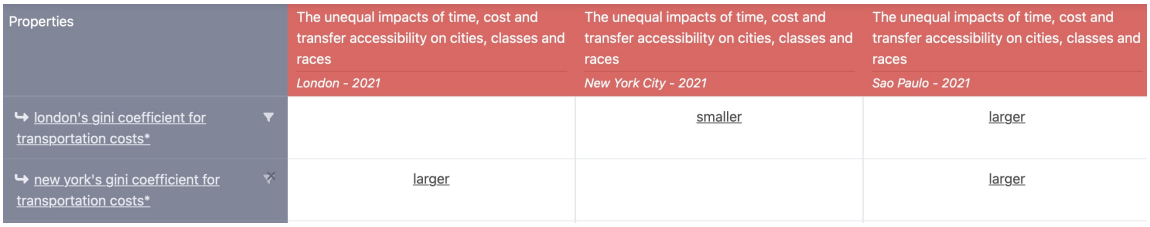


Figure 4.1: Example ORKG comparison with three contributions and two properties

	C1	C2	C3
P1	0	1	1
P2	1	0	1

Table 4.1: Number of values in the comparison

$$pc1 = \frac{0+1+1}{3} = \frac{2}{3} \tag{3}$$

$$pc2 = \frac{1+0+1}{3} = \frac{2}{3} \tag{4}$$

$$\text{average property completeness} = \frac{\text{pc1} + \text{pc2}}{2} \approx 66\%$$

4.1.2 Instance Completeness

Instance completeness is an important metric used to evaluate the coverage and representativeness of a knowledge graph for a given domain. It measures the extent to which all instances of a given class are present in a knowledge graph. Xie [43]

defined instance completeness as the ratio of the number of instances of a class in the knowledge graph to the total number of instances of that class. This means that a high instance completeness score indicates that a knowledge graph contains a large proportion of instances for a given class.

Other researchers have proposed alternative metrics for instance completeness, such as the ratio of missing instances to the total number of instances of a class [38], or the percentage of entities in a class that are correctly identified in the knowledge graph [7]. These alternative metrics may provide a more nuanced assessment of instance completeness, as they take into account not only the presence of instances in a knowledge graph, but also the extent to which these instances are correctly identified or labeled.

Despite these different approaches, instance completeness is generally recognized as a fundamental metric for evaluating the quality and completeness of a knowledge graph. This is particularly important in domains where the identification of all instances of a class is critical, such as in biomedical or scientific domains. Therefore, instance completeness is often used in combination with other metrics, such as property completeness, to provide a more comprehensive evaluation of the quality and coverage of a knowledge graph.

Instance Completeness in the ORKG

According to the Knowledge Graph Maturity Model (KGMM) [20], instance completeness is an essential metric for ensuring the meaningfulness of data in a knowledge graph (KG). The KGMM recommends that all properties and instances in a KG should be complete to satisfy the needs of data consumers. In the KGMM it is suggested that required instances can be determined through various means, such as a representative sample of classes, user evaluation, or a gold standard. In the ORKG realization of the KGMM, this completeness is addressed through predefined property ranges that restrict the data added to the KG within specified ranges, ensuring semantic accuracy at a shallow level. Additional measures, such as human validation and curation of entered data, are also implemented to ensure semantic accuracy.

Overall, the KGMM emphasizes the importance of instance completeness for ensuring meaningful data in a KG, and the ORKG provides tools and measures to help achieve this completeness.

To measure instance completeness in the ORKG, I use a ratio-based approach that considers the total number of literals and resources for a property in a given comparison as the numerator. The denominator is calculated by multiplying the number of distinct literals and resources for the same property in the comparison

with the number of contributions.

To calculate the instance completeness of a property in ORKG, I consider both the number of instances present and their uniqueness across contributions. The metric's numerator represents the total number of instances for a property, while the denominator reflects the product of the number of distinct instances and resources and the number of contributions for that property. A larger denominator indicates a greater level of uniqueness among contributions, whereas a higher numerator suggests a greater amount of instances for that property.

The formulas below outline the calculation process for measuring instance completeness using the previously mentioned criteria:

Equation 5 shows the basic formula for calculating the instance completeness of each property, where $P(C)$ is all values (i.e. all resources and literals) for the property and C is contribution. $|P(C)|$ provides the number of all entries in a cell. To find the average instance completeness across all properties in the comparison, the mean if calculated as shown in Equation 6.

$$\frac{\sum_C |P(C)|}{|\{p : \exists C : p \in P(C)\}| \cdot |C|} \tag{5}$$

$$\frac{\sum_P \frac{\sum_C |P(C)|}{|\{p:\exists C:p\in P(C)\}|\cdot|C|}}{|P|} \tag{6}$$

This metric calculates the information diversity of each property.

An example of a comparison from ORKG can be used to illustrate how to compute the instance completeness metric.

Properties	The unequal impacts of time, cost and transfer accessibility on cities, classes and races *London - 2021*	The unequal impacts of time, cost and transfer accessibility on cities, classes and races *New York City - 2021*	The unequal impacts of time, cost and transfer accessibility on cities, classes and races *Sao Paulo - 2021*
↪ london's gini coefficient for transportation costs*		smaller	larger
↪ new york's gini coefficient for transportation costs*	larger		larger

Figure 4.2: Example ORKG comparison with three contributions and two properties

	C1	C2	C3
P1	0	1	1
P2	1	0	1

Table 4.2: Number of values in the comparison

$$\text{instance completeness p1} = \frac{\#(\text{smaller,larger})}{\#(\text{smaller, larger}) \times \#(\text{C1, C2, C3})} = \frac{2}{6} \tag{7}$$

$$\text{instance completeness p2} = \frac{\#(\text{larger,larger})}{\#(\text{larger}) \times \#(\text{C1, C2, C3})} = \frac{2}{3} \tag{8}$$

$$\text{average instance completeness} = 50\%$$

4.1.3 Population Completeness

Population completeness is a metric used to evaluate the extent to which a knowledge graph (KG) represents all the entities that exist in a given domain. It measures the coverage of a KG in terms of the total number of entities in the domain, and has become an important measure of the quality of a KG. One common approach to measuring population completeness is to compare the entities in the KG with external datasets, such as online databases or registries. For example, Liu et al. [24] used a variety of external sources to evaluate the population completeness of a

biomedical KG. They found that their KG had a high level of completeness for some entities, but lower completeness for others.

Another approach to measuring population completeness is to use statistical sampling techniques to estimate the size and distribution of the entity population. For example, Stadler et al. [35] proposed a method for estimating the size of the population of genes and proteins in a KG using a combination of sampling and extrapolation techniques. Their method involved selecting a representative subset of entities from the KG and extrapolating to estimate the size of the entire population. They found that their method was effective at estimating the size of the population, and could be used to assess the completeness of a KG.

In addition to external sources and statistical sampling, some researchers have proposed other methods for measuring population completeness. For example, Tramesberger et al. [37] proposed a method based on entity co-occurrence patterns. Their method involved analyzing the frequency with which entities appeared together in text, and using this information to estimate the size and distribution of the entity population. They found that their method was effective at identifying missing entities in a KG, and could be used to improve its completeness.

Other researchers have proposed alternative metrics for population completeness that take into account the relationships between entities in the knowledge graph. For example, the Gene Ontology Consortium has developed a metric called coverage, which measures the proportion of annotations in the knowledge graph that are associated with a particular gene [5]. This metric takes into account not only the presence of a gene in the knowledge graph, but also the completeness of the annotations associated with that gene.

Measuring population completeness remains a complex and challenging task due to several factors. One of the main challenges is defining the relevant population for a given domain, which can vary depending on the specific research question. For example, in healthcare, the relevant population may be defined based on a disease state or demographic characteristics [14]. Another challenge is determining the completeness of the reference set used for comparison, such as a gold standard or other benchmark. This can be influenced by various factors, including the availability of data and the expertise of the curators who create the reference set [28]. Overall, accurately measuring population completeness requires careful consideration of these and other factors that may impact the quality and completeness of the data.

Population Completeness in ORKG

According to the KGMM [20], population completeness is defined as the fraction of entities to all other objects in the world. In the context of the ORKG, population completeness is enforced through cardinality restrictions on the property level within templates, which restrict the number of values that can be associated with a given property. This ensures that only the specified cardinality is accepted as input.

Furthermore, KGMM notes that population completeness is closely related to property and instance completeness. In other words, the completeness of the population of entities in a knowledge graph is directly related to the completeness of the properties and instances associated with those entities. This highlights the importance of considering all three aspects of completeness when evaluating the quality and maturity of a knowledge graph.

To calculate population completeness, I employed a method that involves calculating the fraction of literals and resources for a given property in a contribution to a knowledge graph to the total number of literals and resources for that same property in a comparison knowledge graph. Unlike some other methods, this approach does not involve averaging the completeness scores across all properties, but rather calculates a fraction for each individual property. The numerator of this fraction represents the number of literals and resources for the given property in the contribution, while the denominator represents the total number of literals and resources for that same property in the comparison. This method allows for a more fine-grained analysis of population completeness, as it can identify specific properties that may be lacking in completeness.

The formulas used to measure population completeness are described below.

Begin by creating a matrix which is the sum of all elements in each cell of the comparison, as defined in Equation 9, resulting in Matrix X.

$$f(A_j) = \begin{cases} 1, & \text{if there is an element in a cell} \\ 0, & \text{otherwise} \end{cases}$$

$$x_{ij} = \sum_{l=1}^{k} (f(A_j)) \tag{9}$$

$$X = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1n} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2n} \\ \dots & \dots & \dots & \dots & \dots \\ x_{m1} & x_{m2} & x_{m3} & \dots & x_{mn} \end{bmatrix}$$

Calculate the sum of each row in Matrix X using Equation 10, then populate Matrix Y with each value from Matrix X divided by P_r.

$$P_r = \sum_{j=1}^{n} (x_{rj}), r = 1, m \tag{10}$$

$$Y = \begin{bmatrix} \frac{x_{11}}{P_1} & \frac{x_{12}}{P_1} & \frac{x_{13}}{P_1} & \dots & \frac{x_{1n}}{P_1} \\ \frac{x_{21}}{P_2} & \frac{x_{22}}{P_2} & \frac{x_{23}}{P_2} & \dots & \frac{x_{2n}}{P_2} \\ \dots & \dots & \dots & \dots & \dots \\ \frac{x_{m1}}{P_m} & \frac{x_{m2}}{P_m} & \frac{x_{m3}}{P_m} & \dots & \frac{x_{mn}}{P_m} \end{bmatrix}$$

The calculation of population completeness on an ORKG comparison is provided below.

Properties	The unequal impacts of time, cost and transfer accessibility on cities, classes and races *London - 2021*	The unequal impacts of time, cost and transfer accessibility on cities, classes and races *New York City - 2021*	The unequal impacts of time, cost and transfer accessibility on cities, classes and races *Sao Paulo - 2021*
↳ london's gini coefficient for transportation costs*		smaller	larger
↳ new york's gini coefficient for transportation costs*	larger		larger

Figure 4.3: Example ORKG comparison with three contributions and two properties

Measuring completeness in a knowledge graph involves assessing how well the graph represents the real-world entities and relationships it is meant to capture. Various facets of completeness have been explored, including property completeness, instance completeness, and population completeness. Each metric provides valuable

	C1	C2	C3
P1	0	1	1
P2	1	0	1

Table 4.3: Number of values in the comparison

	C1	C2	C3
P1	0	1/3	1/3
P2	1/3	0	1/3

Table 4.4: Example population completeness in the comparison

insights into the quality and comprehensiveness of the data, but also has its own limitations and challenges. For example, measuring population completeness can be complex due to variations in the definition of the relevant population and the accuracy of the reference sets used for comparison. Therefore, it is essential to use a variety of methods and tools to ensure the completeness of the knowledge graph.

4.2 Accuracy

Accuracy is a critical aspect of any knowledge graph, as it determines the reliability of the information it contains. In the context of knowledge graphs, accuracy refers to the extent to which the information represented in the graph corresponds to the real-world entities and relationships it is intended to capture [30]. A knowledge graph that is inaccurate may lead to incorrect conclusions, and can also impact downstream applications that rely on the data.

There are several challenges associated with ensuring accuracy in a knowledge graph. One of the main challenges is the sheer volume and complexity of the data. Knowledge graphs often contain large amounts of data from a variety of sources, which can be difficult to reconcile and integrate accurately [1]. Another challenge is the need to deal with noisy or incomplete data, which can result in errors and inaccuracies in the graph [30].

To address these challenges, researchers have developed various methods and tools to ensure the accuracy of knowledge graphs. One common approach is to use crowdsourcing to validate and verify the data in the graph. This involves enlisting the help of a large number of people to manually review and correct the data, which can improve the accuracy of the graph [22]. Another approach is to use automated techniques such as machine learning to identify and correct errors in the graph [30].

In addition to these techniques, there are also metrics and benchmarks that can be used to evaluate the accuracy of a knowledge graph. One such metric is precision and recall, which measures the percentage of correct answers given by the graph as compared to a set of ground truth data [27]. Another metric is the F1 score, which combines precision and recall into a single measure [30]. These metrics can help identify areas where the graph is inaccurate and guide efforts to improve its accuracy.

In conclusion, accuracy is a critical aspect of any knowledge graph, and ensuring its accuracy requires a variety of methods, tools, and metrics. While there are many challenges associated with ensuring accuracy, ongoing efforts to improve the accuracy of knowledge graphs are essential for their continued use and usefulness.

4.2.1 Semantic Accuracy

Semantic accuracy is an important aspect of knowledge graphs that determines their usefulness in various applications. In simple terms, semantic accuracy refers to how well the knowledge graph represents the intended meaning of the concepts and relationships in the real world. This means that the graph should be able to capture not only the syntactic or structural aspects of the data but also the underlying semantics [30]. Semantic accuracy can be evaluated using various measures such as precision, recall, F1 score, and semantic distance [11]. These measures assess how well the knowledge graph aligns with the expected semantics based on external sources, such as domain ontologies, taxonomies, or expert knowledge.

One of the primary challenges in achieving semantic accuracy is the heterogeneity of the data sources used to construct the knowledge graph. The data sources may use different terminologies, definitions, or even implicit assumptions, leading to inconsistencies or conflicts in the semantics of the data [1]. Therefore, it is crucial to reconcile the semantic heterogeneity by aligning the data to a common vocabulary or ontology, such as the Linked Open Data (LOD)[22] or the Web Ontology Language (OWL)[27]. This alignment can be achieved through various methods such as ontology matching, entity linking, or named entity recognition.

Another challenge in achieving semantic accuracy is the scalability and automa-

tion of the evaluation process. Manual evaluation by experts can be time-consuming and costly, especially for large and dynamic knowledge graphs. Therefore, several automatic evaluation methods have been proposed, such as the use of reference datasets or crowdsourcing platforms [11]. These methods can provide a benchmark for semantic accuracy and allow for continuous evaluation and improvement of the knowledge graph.

In conclusion, achieving semantic accuracy in knowledge graphs is essential for their effectiveness and usability. It requires addressing the challenges of semantic heterogeneity, scalability, and automation in the construction and evaluation process. Therefore, it is essential to use a combination of approaches, including alignment to common vocabularies, automatic evaluation methods, and expert input, to ensure the semantic accuracy of the knowledge graph [30, 11].

Semantic Accuracy in ORKG

Semantic accuracy is a crucial quality measure for knowledge graphs and data organization systems. As defined by the KGMM [20], it is the degree to which data values accurately depict phenomena in the real world. In the ORKG realization, semantic accuracy is ensured on a shallow level through predefined property ranges that limit data entries to specific ranges. Additionally, human validation and curation of entered data provide further efforts to guarantee semantic accuracy. From a data consumer's perspective, semantic accuracy enables the construction of meaning from the data. For example, if the director's name for a movie is inaccurate, the data consumer will be unable to utilize the data effectively, rendering it incomplete and unbeneficial.

To measure the semantic accuracy of the ORKG at the contribution level, I focused on the use of templates in contributions. Templates define a structured representation of the data that can be entered in a contribution.Using a template for every contribution in a comparison ensures semantic accuracy.

Contributions in ORKG are associated with classes, and any additional contribution-level classes are presumed to be template classes. Thus, I can check what percentage of contributions in a comparison share a class assignment other than the default classes of *contribution* and *resource*. The overall semantic accuracy of a comparison can be calculated by dividing the number of contributions T with the same class other than *contribution* and *resource* by the total number of contributions C, as shown in Equation 11.

$$T = \{c \in C : \text{class } k : c \text{ is instance of } k \notin \{\text{contribution, resource }\}\}$$

$$\frac{|T|}{|C|} \tag{11}$$

The following computation of semantic accuracy on an ORKG comparison is shown.

Properties	The unequal impacts of time, cost and transfer accessibility on cities, classes and races *London - 2021*	The unequal impacts of time, cost and transfer accessibility on cities, classes and races *New York City - 2021*	The unequal impacts of time, cost and transfer accessibility on cities, classes and races *Sao Paulo - 2021*
↪ london's gini coefficient for transportation costs*		smaller	larger
↪ new york's gini coefficient for transportation costs*	larger		larger

Figure 4.4: Example ORKG comparison with three contributions and two properties

Instance of: Results, R288296, Contribution

Figure 4.5: Classes in the first contribution of the comparison in contribution's level

Instance of: Results, R288296, Contribution

Figure 4.6: Classes in the second contribution of the comparison in contribution's level

Instance of: Results, R288296, Contribution

Figure 4.7: Classes in the third contribution of the comparison in contribution's level

In this Comparison, there are three contributions, each containing a template and class ID other than "contribution" and "resource". Upon inspection, I find that the template and class ID for the contributions are the same. This indicates that there is 100% semantic accuracy at the contribution level.

Semantic accuracy is a critical aspect of data quality, and it refers to the degree to which data values accurately reflect real-world phenomena. While several approaches have been proposed to measure semantic accuracy, it is still a challenging task due to the subjective nature of meaning and the complexity of modeling the real world. As Garcıa-Silva et al. [11] explain, "Semantic accuracy is an ongoing challenge, as it requires not only the selection of appropriate data sources but also their integration and interpretation, taking into account the inherent ambiguity of natural language and the diversity of perspectives and needs of the various stakeholders involved". Therefore, ensuring semantic accuracy requires not only automated techniques but also human validation and curation to guarantee that data is meaningful and useful.

4.3 Implementation

To compute the measures of property completeness, instance completeness, population completeness, and semantic accuracy, I utilize Python and the Pandas library. The code is implemented in Jupyter Notebook, allowing me to document my analysis and share it with others.

The input for analysis is the Resource Id of comparisons, which I used to extract the necessary data for computations. The output of analysis is the average of the computed measures, providing a summary of the quality and accuracy of the knowledge graph data.

I utilize the SPARQL query language to retrieve the necessary data for analysis. SPARQL is a standard query language used to retrieve and manipulate data from RDF (Resource Description Framework) databases, which are commonly used to store knowledge graphs. Using SPARQL, I extract the required information from the comparisons, including the property, contribution and instance values, and use them to compute the measures of completeness and accuracy.

In my analysis, I use various Python packages, including Pandas [26], to efficiently modify and analyze the data. Other packages, such as NumPy [12] and Matplotlib [18], are also used to perform statistical analysis and visualize the results.

The codes I used in my analysis are available on my GitHub repository[1].

[1] `https://github.com/Atiyeh-MKH/analysis_process_for_quality_of_KG`

Chapter 5

Evaluation

In order to assess the quality of the comparisons in the ORKG, I conduct an evaluation of nearly 400 comparisons from two curation grants. The analysis focuses on four key measures: property completeness, instance completeness, population completeness and semantic accuracy.

To evaluate the quality of the comparisons, I use a combination of manual and automated methods. I also considered comparisons with different versions and histories to assess the quality over time.

To present my findings, I provide a series of tables and charts that summarize the quality measurements for each comparison, as well as the breakdown of the measurements by component and over time.

In addition, I would like to mention that due to the extensive nature of the analysis, it was not possible to include all tables and charts in the thesis. Therefore, I have selected representative examples for discussion in the thesis, while the complete set of tables and charts is available on my GitHub repository[1]. Readers interested in exploring the details of the analysis can access the scripts and data on the repository.

5.1 Evaluation of Quality Measures

In this section, I describe the quality measures used to evaluate the comparisons in the ORKG curation grants. These quality measures include completeness of both properties and instances, as well as population completeness and semantic accuracy.

Completeness measures the extent to which a comparison includes all relevant properties and instances. For example, a comparison that lacks important properties

[1] `https://github.com/Atiyeh-MKH/analysis_process_for_quality_of_KG`

or instances may have lower completeness scores than one that includes all relevant information. Coefficient values are assigned to each completeness measure to give them a specific weight in the overall evaluation of the comparison.

Population completeness is a measure of the extent to which a comparison includes all relevant entities from the domain of interest. In this evaluation, I calculate population completeness for two selected comparisons with histories and analyzed how it change over time.

Semantic accuracy was measured by calculating what percentage of contributions in a comparison use the templates. This is done to ensure that the contributions adhere to the standard structure and language used in the ORKG, which helps to ensure consistency and compatibility between different contributions.

5.2 Analysis of ORKG Curation Grants

In this section, I present the results of the evaluation of the ORKG curation grants using the quality measures described in the previous section. For the evaluation, I randomly select 12 comparisons for each curation grant. For the first ORKG curation grant, I choose two comparisons for each month from July to December 2021. Similarly, for the second ORKG curation grant, I choose two comparisons for each month from June to November 2022.

The purpose of selecting a small subset of comparisons is to evaluate how the quality measures would perform and to understand how the results would appear before applying the measures to the entire ORKG.

I calculate the quality of each comparison using an average method, in which I give a coefficient of 2 to both property completeness and instance completeness, and add one-tenth of the value of semantic accuracy to it. I then divide the sum of these values by 4. I choose this average method because a usual average would not have provided a clear result. Additionally, I only measure semantic accuracy at the contribution level although it is possible for semantic accuracy to exist at different levels of depth within the knowledge graph. Furthermore, there were numerous instances of zero for semantic accuracy and because of that I use one-tenth of the value of this measure and it affects the overall average values if it is not zero. I also do not consider population completeness in the quality average.

The data displayed below provides a clear depiction of the outcomes.

1st ORKG Curation Grant, July-December 2021					
Resource Id	Date	Property Compeletness in percent	Instance Compeletness in percent	Semantic Accuracy in percent	Avg Quality in percent
R135715	Jul 21	100,00	49,07	0,00	74,54
R138179	Jul 21	81,61	32,06	0,00	56,84
R139357	Aug 21	100,00	41,61	25,00	71,43
R139295	Aug 21	100,00	37,17	0,00	68,59
R141783	Sep 21	95,45	45,43	0,00	70,44
R143856	Sep 21	85,71	31,12	0,00	58,42
R144936	Oct 21	100,00	34,03	0,00	67,02
R148383	Oct 21	100,00	29,63	0,00	64,82
R151435	Nov 21	100,00	30,23	0,00	65,12
R148678	Nov 21	100,00	37,04	100,00	71,02
R160527	Dec 21	97,22	31,59	0,00	64,41
R160597	Dec 21	100,00	23,85	0,00	61,93
Avg		96,67	35,24	10,42	66,21

Figure 5.1: Average Quality in Random Comparisons - 1st Curation Grant

2nd ORKG Curation Grant, June-November 2022					
Resource Id	Date	Property Compeletness in percent	Instance Compeletness in percent	Semantic Accuracy in percent	Avg Quality in percent
R193278	Jun 22	85,71	44,44	0,00	65,08
R193505	Jun 22	99,76	29,15	0,00	64,46
R201854	Jul 22	25,25	42,52	0,00	33,89
R202362	Jul 22	97,91	42,90	0,00	70,41
R210531	Aug 22	47,69	42,79	0,00	45,24
R211935	Aug 22	12,67	46,93	100,00	32,30
R217515	Sep 22	90,90	13,45	100,00	54,68
R219443	Sep 22	51,50	43,31	0,00	47,41
R224810	Oct 22	100,00	29,58	7,69	64,98
R233016	Oct 22	36,05	49,54	100,00	45,30
R257000	Nov 22	36,00	44,00	0,00	40,00
R259184	Nov 22	58,88	36,38	100,00	50,13
Avg		61,86	38,75	33,97	51,15

Figure 5.2: Average Quality in Random Comparisons - 2nd Curation Grant

It is surprising to see a drop in quality between the two curation grants, especially given my expectations for the ORKG's quality to improve over time. The tables and chart 5.1, 5.2, and 5.3 clearly indicate the average quality measures for the randomly selected comparisons in both grants, with the second grant demonstrating a reduction compared to the first.

It is crucial to remember that these comparisons represent only a small subset of the full ORKG, and more research is required to make definitive conclusions about the overall quality of the knowledge graph.

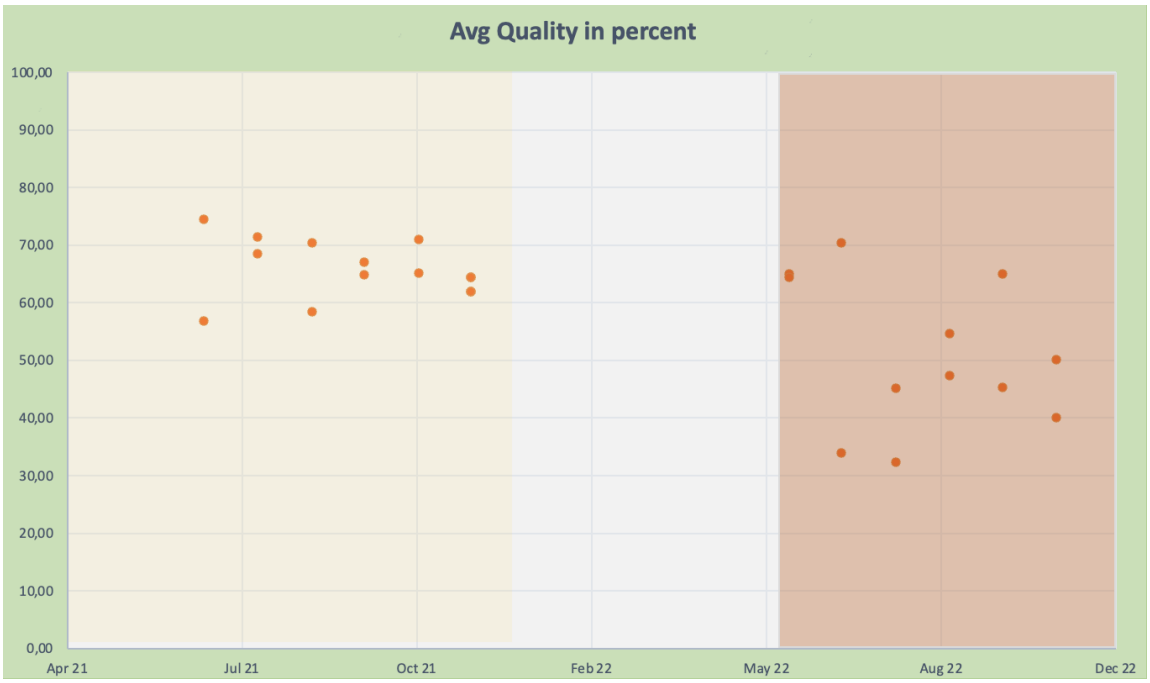


Figure 5.3: Average Quality of Randomly Selected Comparisons in the 1st and 2nd ORKG Curation Grants

In the next step, I aim to measure the average quality for all the comparisons in the first and second ORKG curation grant in exactly the same order that is in the ORKG, and group them by the name of contributors to have a better and more accurate understanding of the results. I use the same average method mentioned earlier to calculate the quality of each comparison.

Table 5.4 displays the average quality scores of a selected subset of comparisons completed by a single group among 21 ORKG groups in the first and second curation grants. Each group contains a set of comparisons, and Table 5.4 presents the average quality scores of this partial set of comparisons for the selected group.

Resource Id	Date	Property Compeletness in percent	Instance Compeletness in percent	Semantic Accuracy in percent	Avg Quality in percent
Digital City Planner					
R139268	Aug 21	89,81	19,75	0,00	**54,78**
R157384	Dec 21	48,37	33,59	0,00	**40,98**
R157531	Dec 21	88,88	20,25	0,00	**54,57**
R159000	Dec 21	88,88	20,25	0,00	**54,57**
R215443	Sep 22	88,23	20,46	0,00	**54,35**
R137510	Jul 21	100,00	100,00	0,00	**100,00**
R139568	Aug 21	100,00	100,00	0,00	**100,00**
R74314	Apr 21	100,00	51,11	0,00	**75,56**
R78489	May 21	100,00	51,33	0,00	**75,67**
R135668	Jul 21	27,17	41,84	0,00	**34,51**
R135920	Jul 21	100,00	50,34	0,00	**75,17**
R139567	Aug 21	90,78	36,11	0,00	**63,45**
R141627	Sep 21	68,00	45,33	0,00	**56,67**
R141628	Sep 21	100,00	35,71	0,00	**67,86**
R141845	Sep 21	100,00	35,71	0,00	**67,86**

Figure 5.4: Average Quality of Partial Comparisons in Selected ORKG Group

To provide a visual representation of the quality scores for the selected group, a box plot chart 5.5is included in Table5.4.

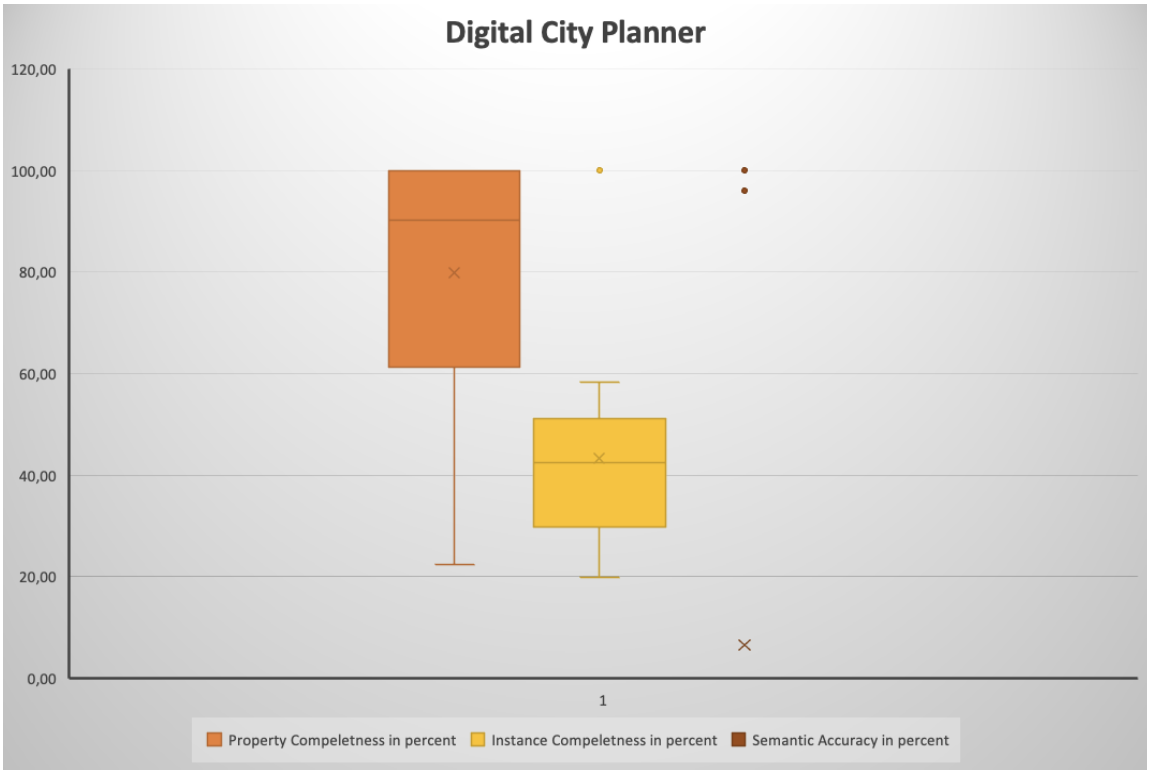


Figure 5.5: Average Quality of Comparisons in Selected ORKG Group

After measuring the average quality for almost 400 comparisons in the first and second ORKG curation grants, I found a diverse range of results in each quality measure. To better understand these results and obtain more information, I focus on each characteristic individually. This approach allows me to gain insights into the strengths and weaknesses of the ORKG in terms of property completeness, instance completeness, and semantic accuracy. By analyzing the data in this way, I gain a more comprehensive understanding of the quality of the knowledge graph and identify areas for improvement.

I start with property completeness and order the comparisons by month, calculating the average property completeness for each month from July 2021 to December 2022. To visualize these results, I create two separate charts: chart 5.6 displays the property completeness for each month, while chart 5.7 depicts the average property completeness for each month during the same period.

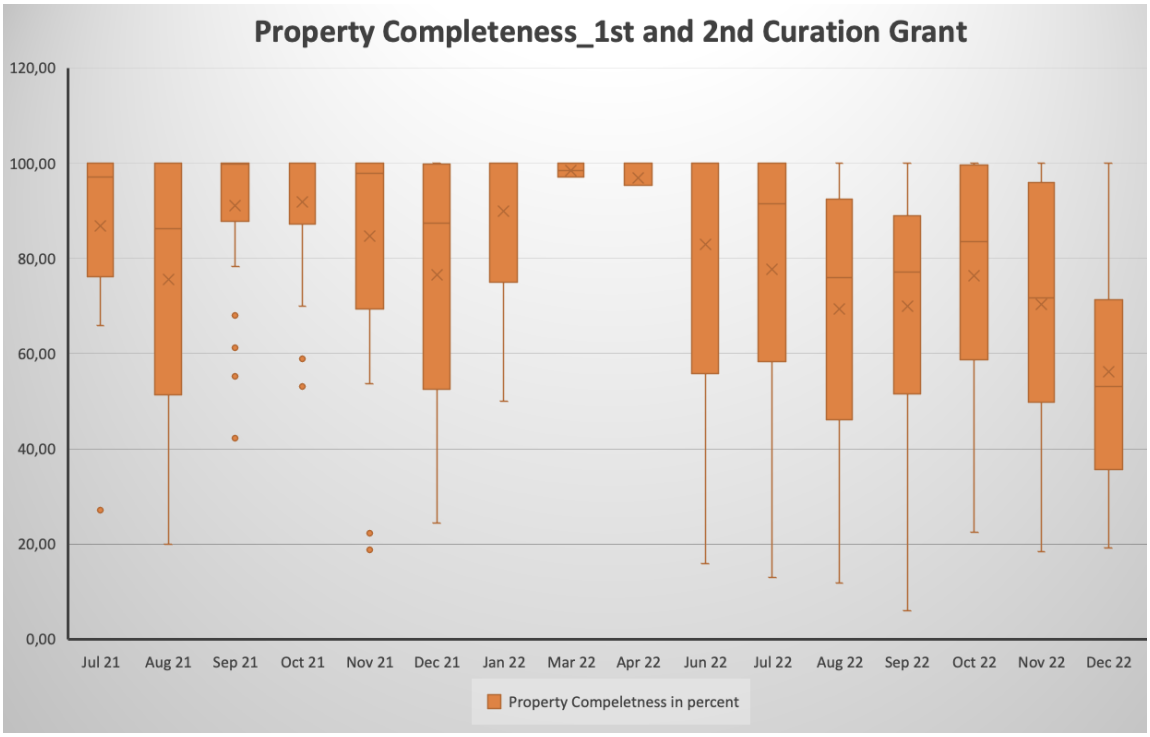


Figure 5.6: Property Completeness (July 2021 - December 2022)

Figure 5.7: Average Property Completeness (July 2021 - December 2022)

The average property completeness for the ORKG curation grants varies significantly across different months. For instance, the months of September and October 2021 have the highest average property completeness, while December 2022 has the lowest average property completeness. However, it's essential to note that the number of comparisons in each month is different, which may have affected the overall results. Therefore, to gain a more comprehensive understanding of the property completeness's distribution across different months, it's better to use the box plot.

Instance completeness is another important quality measure for a knowledge graph, and I also evaluate this measure for the ORKG curation grants during the same period as property completeness. Similar to property completeness, I group the comparisons by month and calculate the average instance completeness for each month. The results of the evaluation are presented in chart 5.9, and a box plot chart of the instance completeness for each month is shown in chart 5.8.

During the period from July 2021 to December 2022, the average instance completeness varied greatly, as seen in the following values: In January 2022, the average instance completeness was 49.32%, while in March 2022, it dropped to 21.67%. There

was a significant increase in April 2022, with an average instance completeness of 61.45%. However, it fell again in the following months, reaching 36.30% in August 2022. It is important to note that the number of comparisons in different months varied significantly, which can impact the results.

Comparing the average instance completeness with the average property completeness in each month, it can be seen that the scores for instance completeness are generally lower than those for property completeness. While the average property completeness scores remained above 75% in most months, the average instance completeness scores varied greatly, with some months having scores below 30%. This suggests that there may be issues with the completeness of instances in some of the comparisons, which could impact the overall quality of the results. Further investigation may be needed to identify the root causes of these issues and take corrective actions.

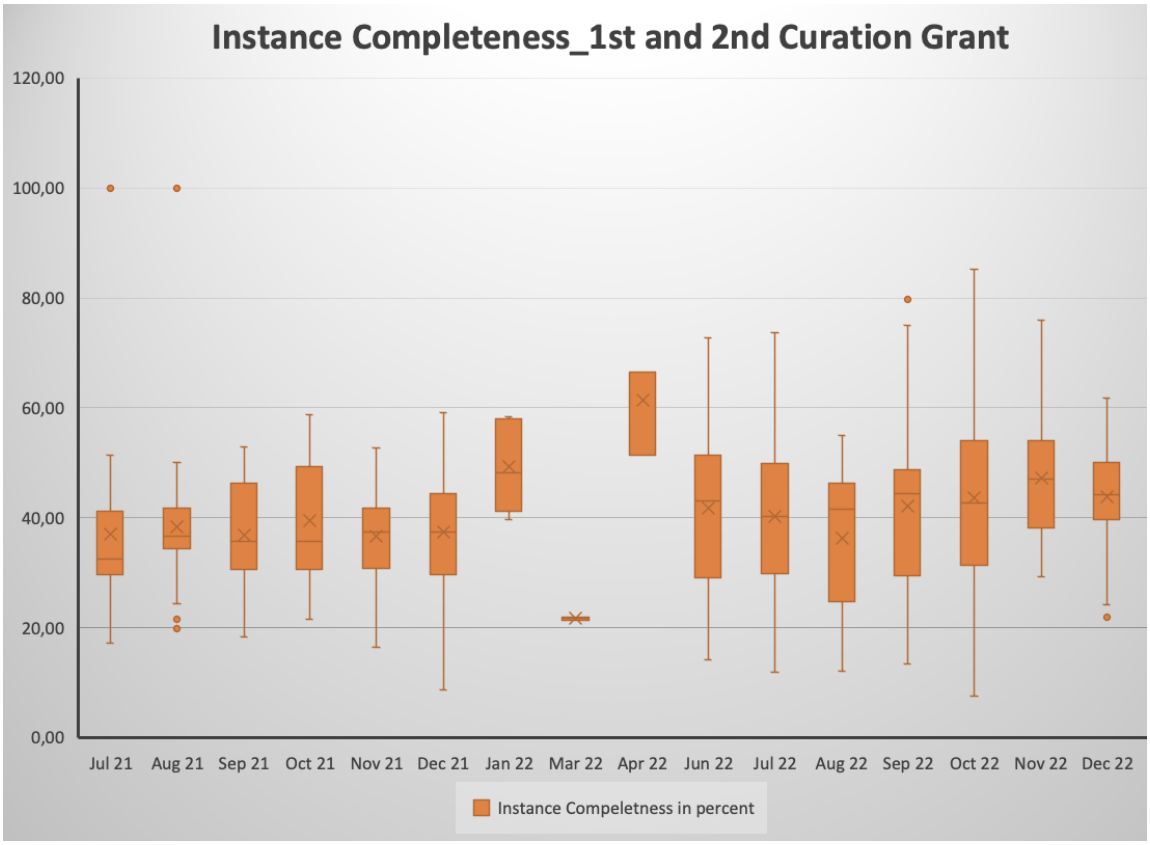


Figure 5.8: Instance Completeness (July 2021 - December 2022)

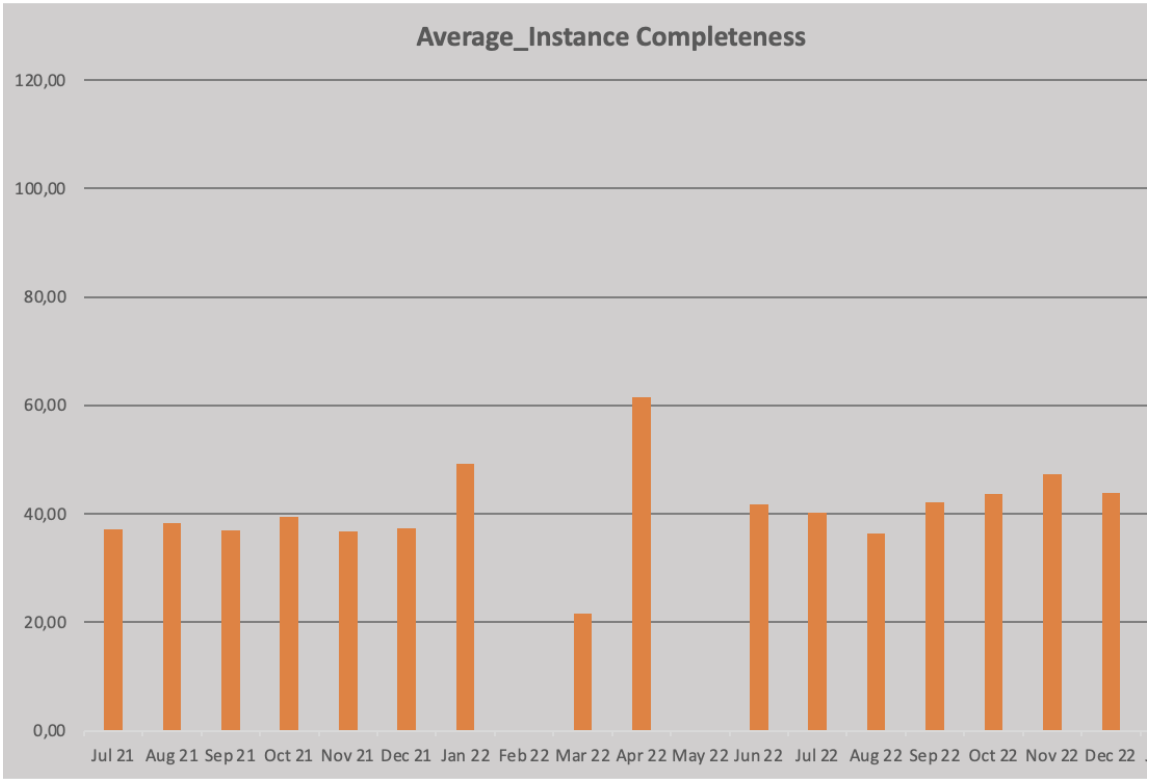


Figure 5.9: Average Instance Completeness (July 2021 - December 2022)

Moving on to the next characteristic, semantic accuracy, I measure the results during the same period from July 2021 till December 2022 for all the comparisons in the first and second ORKG curation grant. The results are shown in the chart5.10 and chart5.11 below.

In analyzing the results for semantic accuracy, it's important to note that using averages might not be the best approach. This is because the values for semantic accuracy are either at zero or 100, with very few comparisons having a value in between. As such, it might be more useful to examine the distribution of values and identify any patterns or trends. To help visualize this, Chart 5.12 has been included, which shows the distribution of semantic accuracy values for all comparisons during the period from July 2021 to December 2022.

Upon examining the chart 5.12, it's clear that there are a significant number of comparisons that have a semantic accuracy value of zero.

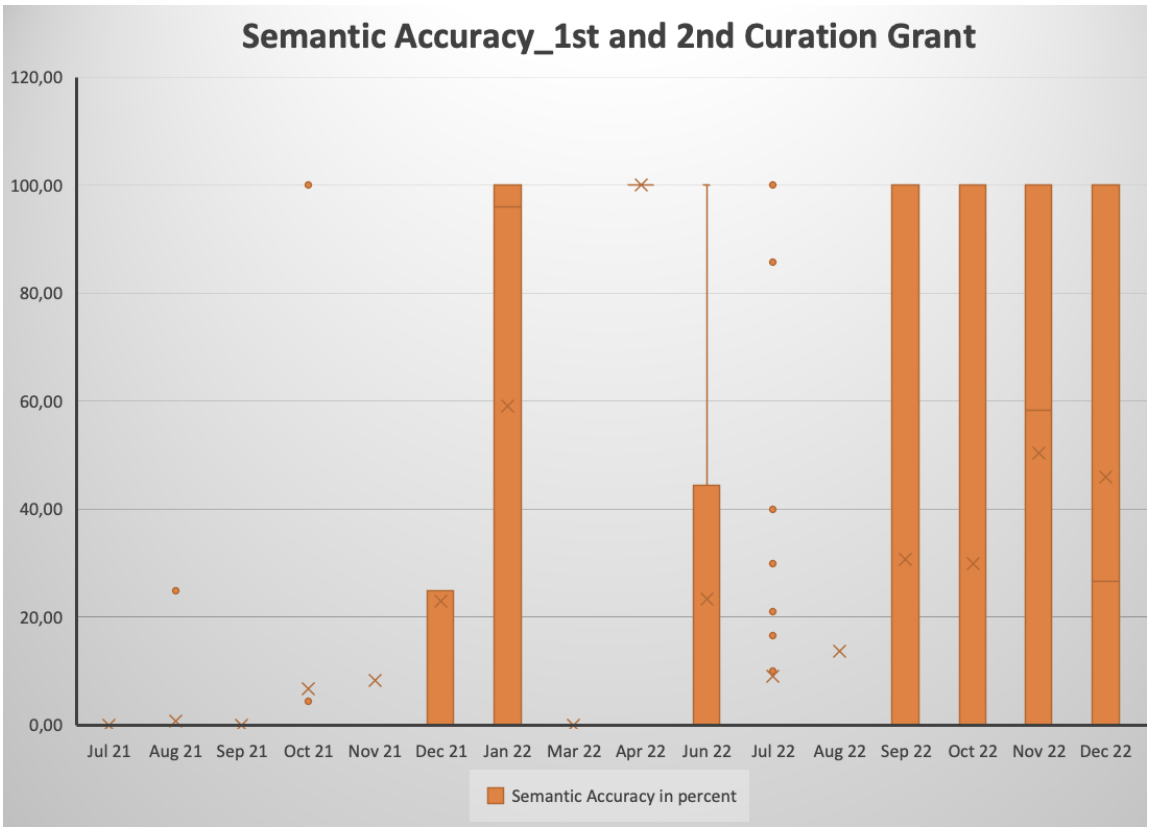


Figure 5.10: Semantic Accuracy (July 2021 - December 2022)

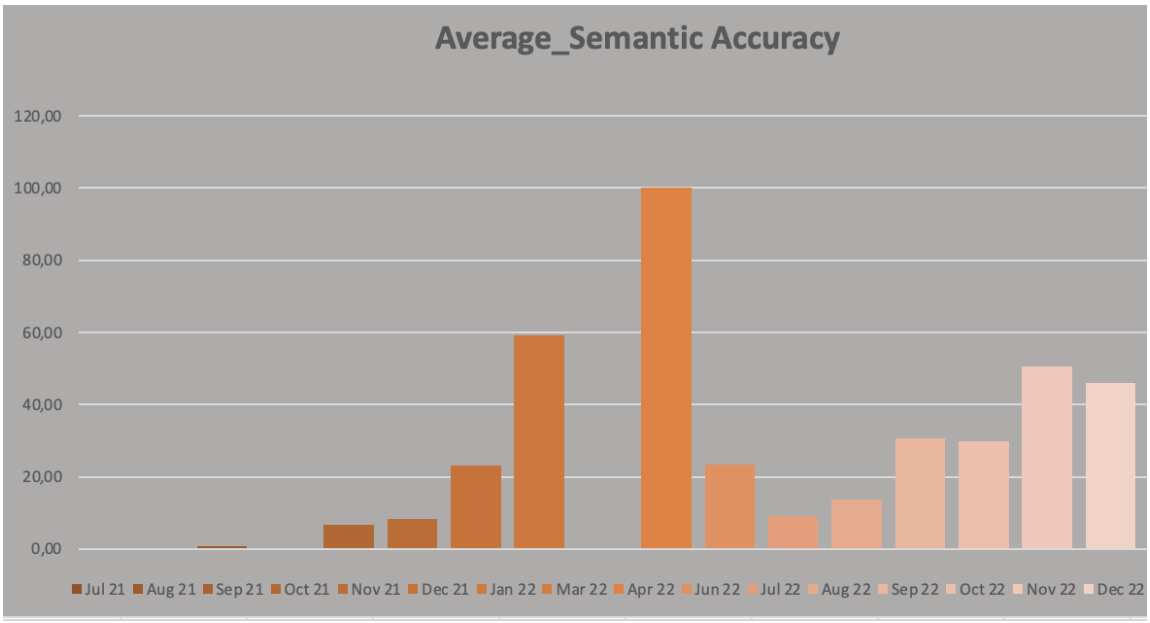


Figure 5.11: Average Semantic Accuracy (July 2021 - December 2022)

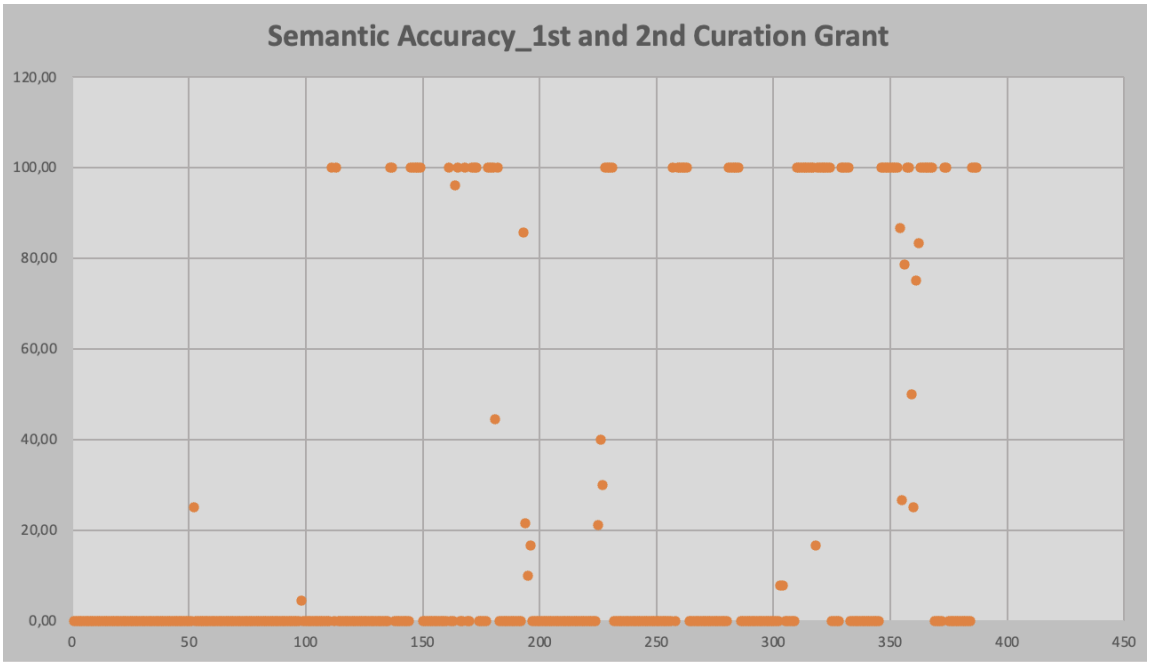


Figure 5.12: Semantic Accuracy (July 2021 - December 2022)

Further investigation reveals that many of these comparisons don't have a corresponding template, highlighting an issue with the ORKG system. In analyzing the semantic accuracy at the contribution level across almost 400 comparisons, this issue has been identified as a major concern that needs to be addressed.

By using templates for all contributions in a comparison, the overall semantic accuracy of the ORKG system could be improved.

While measuring the semantic accuracy of comparisons in the ORKG system, I noticed that there were only a few comparisons that had a between value for this characteristic. This made me curious to investigate further and analyze the 21 comparisons that fell within this range. After examining their templates, I have discovered something unusual in one of the comparisons. The only class used in this comparison was "ContributionDeleted," which means that contributors were able to add empty contributions during the creation of comparisons in ORKG. This is problematic as it can lead to inaccurate or incomplete data being added to the system. By identifying this issue, efforts can be made to improve the system and prevent such mistakes from occurring in the future

5.3 Evaluating Completeness through Different Approaches

In the ORKG, some comparisons have a history, meaning that different versions of the comparison exist, which may include changes such as adding or removing contributions or properties. These comparisons provide an excellent sample for evaluating the quality of the comparison over time, as changes may impact the completeness and other characteristics of the comparison. Therefore, in this subsection, I focus on comparisons with multiple versions extracted from the first and second ORKG curation grant, which includes a total of 30 sets of comparisons.

By analyzing the completeness of each version, I can determine whether changes made to the comparison had an impact on its overall completeness.

Analyzing these comparisons can provide valuable insights into the quality of comparisons over time, and it can help improve the curation process of comparisons in ORKG.

Figure 5.13 shows a portion of the table that displays the average quality for the five versions of one of the comparisons with history in the first ORKG curation grant:

Resource Id	Version	Contributor	Date	Property Compeletness in percent	Instance Compeletness in percent	Semantic Accuracy in percent	Avg Quality in percent
R139268	1	Alena Begler	Aug 21	89,81	19,75	0,00	**54,78**
R157384	2	Alena Begler Natalia Chichcova	Dec 21	48,37	33,59	0,00	**40,98**
R157531	3		Dec 21	88,88	20,25	0,00	**54,57**
R159000	4	Alena Begler	Dec 21	88,88	20,25	0,00	**54,57**
R215443	5		Sep 22	88,23	20,46	0,00	**54,35**
Avg				**80,83**	**22,86**	**0,00**	**51,85**

Figure 5.13: Quality - Five Versions of a Comparison from 1st Curation Grant

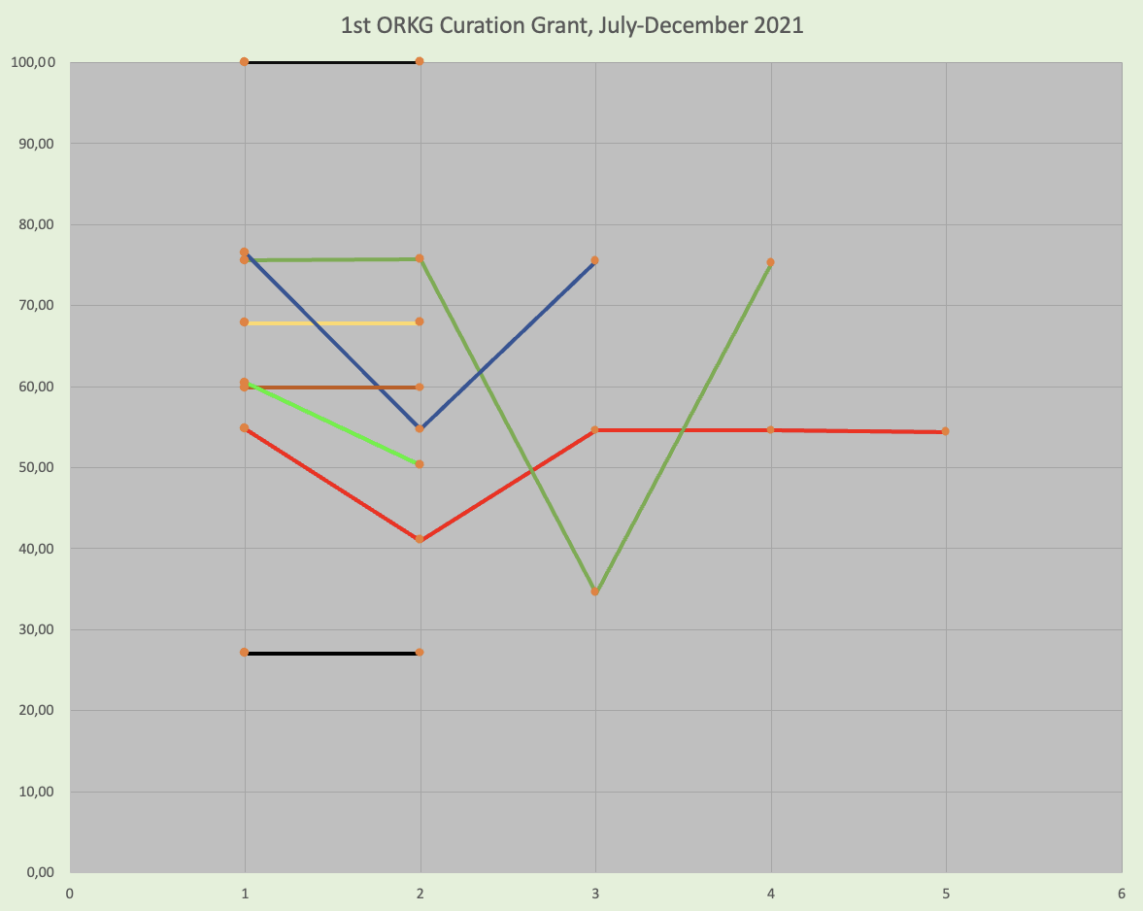


Figure 5.14: Average Quality Changes Across Versions of Comparisons

Additionally, line chart 5.14 displays the average quality of all comparisons with history in the first ORKG curation grant and their changes during the different versions. For instance, the example shown in Table 5.13 is represented by the red line in line chart 5.14, in which the second version of the comparison shows a significant drop in quality compared to the other versions, with a completeness score of 40.98%. This indicates that there may have been some issues with the contributions or properties added in this version. It would be worthwhile to investigate this version in more detail to identify any potential issues and improve the overall quality of the comparison.

These visual aids allow for a more in-depth analysis of how the quality of comparisons with history has evolved and can be used to identify patterns or trends in their development.

Additionally, the example of the comparison with 5 versions shown in figure 5.13 can be used to illustrate changes in population completeness, which are not reflected in the average quality. The population completeness of the fourth and fifth versions of this comparison discussed earlier are presented in Figure 5.15 using separate tables. These tables are formatted similarly to how comparisons are presented in ORKG, providing a user-friendly visualization for users to assess the quality of a comparison in terms of population completeness. By utilizing this approach, users can quickly identify changes in population completeness over time, helping to highlight potential issues or areas of improvement for the comparison.

When looking at a single table, users can easily identify if population completeness is 0, which may indicate a missing value. In such cases, they should check if the value was not inputted or if the paper does not include it. In addition, if a cell has very high property completeness while others for the same predicate do not, users should check if they forgot to include some values in other cells.

In terms of population completeness, the fourth and fifth versions of the comparison 5.15 differ clearly. The fourth version includes 18 properties, while the fifth version only includes 17. To properly compare each property between the two versions, it is necessary to consider properties with the same "property ID". Furthermore, it is important to note that there are 2 instances of zero values in these tables. While this may not be problematic at the moment, it could become an issue as the number of versions for a comparison increases, and the number of zero values accumulates, potentially impacting the overall quality.

Despite this, the tables offer a helpful visualization for ORKG users, enabling them to quickly assess the population completeness of a comparison across different versions.

R159000 4th Version Alena Begler	C.1	C.2	C.3	C.4	C.5	C.6	Property ID
P.1	16,67	16,67	16,67	16,67	16,67	16,67	P32
P.2	16,67	16,67	16,67	16,67	16,67	16,67	P41128
P.3	12,5	25	12,5	25	12,5	12,5	P39139
P.4	16,67	16,67	16,67	16,67	16,67	16,67	P41142
P.5	9,09	63,64	18,18	0	0	9,09	P41146
P.6	14,29	14,29	14,29	14,29	28,57	14,29	P41134
P.7	15,38	15,38	7,69	38,46	7,69	15,38	P41137
P.8	16,67	16,67	16,67	16,67	16,67	16,67	P41144
P.9	11,11	33,33	11,11	11,11	11,11	22,22	P41141
P.10	9,09	9,09	27,27	9,09	27,27	18,18	P41135
P.11	9,52	33,33	9,52	19,05	14,29	14,29	P41145
P.12	10	30	10	20	20	10	P41132
P.13	16,67	16,67	16,67	16,67	16,67	16,67	P41140
P.14	16,67	33,33	8,33	16,67	8,33	16,67	P41138
P.15	6,25	31,25	6,25	43,75	6,25	6,25	P41133
P.16	12,5	12,5	12,5	12,5	25	25	P41131
P.17	12,5	25	12,5	12,5	25	12,5	P41136
P.18	14,29	14,29	14,29	28,57	14,29	14,29	P41139

R215443 5th Version Alena Begler	C.1	C.2	C.3	C.4	C.5	C.6	Property ID
P.1	16,67	16,67	16,67	16,67	16,67	16,67	P41128
P.2	12,5	25	12,5	25	12,5	12,5	P39139
P.3	16,67	16,67	16,67	16,67	16,67	16,67	P41142
P.4	9,09	63,64	18,18	0	0	9,09	P41146
P.5	14,29	14,29	14,29	14,29	28,57	14,29	P41134
P.6	15,38	15,38	7,69	38,46	7,69	15,38	P41137
P.7	16,67	16,67	16,67	16,67	16,67	16,67	P41144
P.8	11,11	33,33	11,11	11,11	11,11	22,22	P41141
P.9	9,09	9,09	27,27	9,09	27,27	18,18	P41135
P.10	9,52	33,33	9,52	19,05	14,29	14,29	P41145
P.11	10	30	10	20	20	10	P41132
P.12	16,67	16,67	16,67	16,67	16,67	16,67	P41140
P.13	16,67	33,33	8,33	16,67	8,33	16,67	P41138
P.14	6,25	31,25	6,25	43,75	6,25	6,25	P41133
P.15	12,5	12,5	12,5	12,5	25	25	P41131
P.16	12,5	25	12,5	12,5	25	12,5	P41136
P.17	14,29	14,29	14,29	28,57	14,29	14,29	P41139

Figure 5.15: Population Completeness in 4th and 5th Version of a Comparison

In addition to the analysis of the changes in the quality of comparisons with different versions, the data extracted from the first and second ORKG curation grants allowed for a comparison of the same contributor's work in both grants, specifically in the area of comparisons with history. This consideration provides insights into whether having more experience working on ORKG has any impact on the quality of comparisons created by the same person.

Figure 5.16 shows a table of comparisons created by a single contributor who participated in both grants but created two sets of comparisons with history only in the second ORKG curation grant, allowing for an analysis of the changes in the quality of his work over time.

Resource Id	Version	Contributor	Date	Property Compeletness in percent	Instance Compeletness in percent	Semantic Accuracy in percent	Avg Quality in percent
R191023	1	Enrique Iglesias	Jun 22	100,00	59,98	100,00	**82,49**
R191975	2		Jun 22	100,00	59,98	100,00	**82,49**
Avg				**100,00**	**59,98**	**100,00**	**82,49**
R185262	1	Enrique Iglesias	Apr 22	95,47	66,44	100,00	**83,46**
R185267	2		Apr 22	100,00	51,48	100,00	**78,24**
R191658	3		Jun 22	100,00	51,48	100,00	**78,24**
R191978	4		Jun 22	100,00	72,76	100,00	**88,88**
Avg				**98,87**	**60,54**	**100,00**	**82,20**

Figure 5.16: Quality of Comparisons with the Same Contributor

Additionally, a line chart of the same contributor's comparison sets over time is presented in figure 5.17."

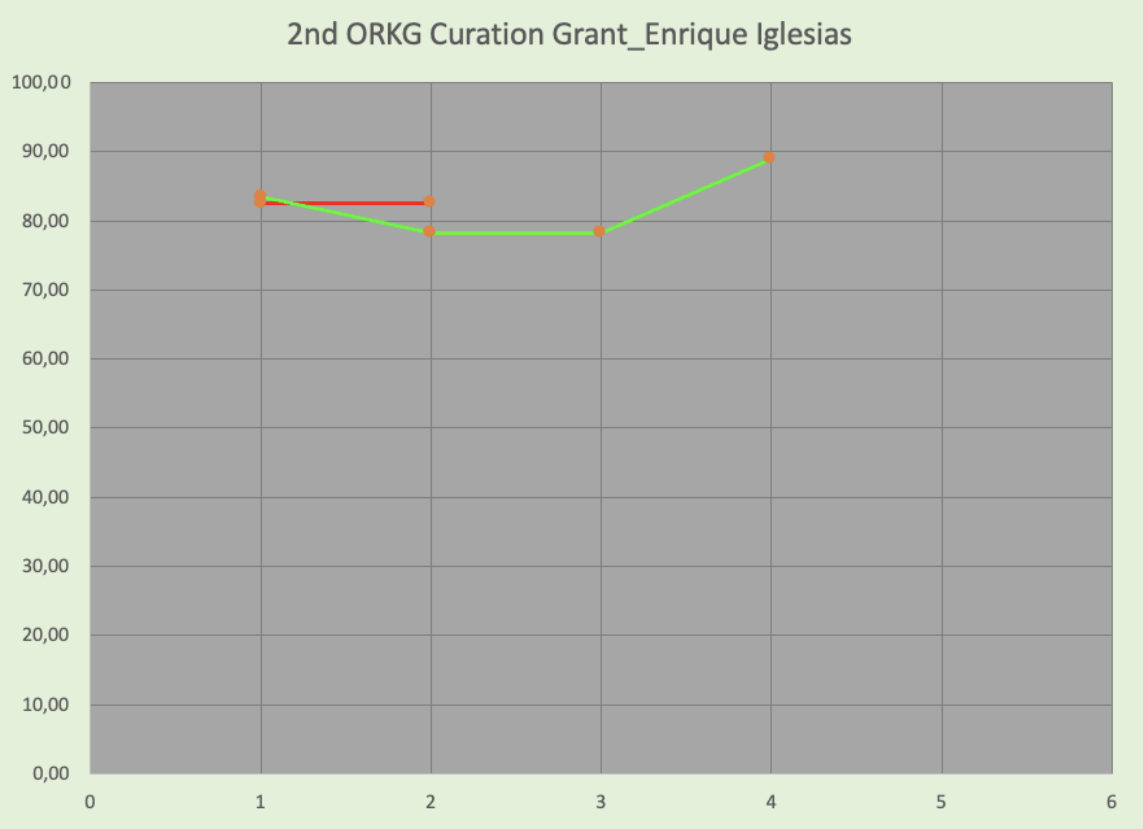


Figure 5.17: Quality of Comparisons with the Same Contributor

Overall, it is difficult to draw any firm conclusions about the contributor's improvement over time based on these limited data points.

It seems that the contributor in question has consistently produced high-quality comparisons, especially in the areas of property completeness and semantic accuracy. However, there is some variability in instance completeness, which may warrant further attention. Overall, the quality of the second set of comparisons with four versions (green line 5.17) appears to be slightly higher than the first set with two

versions (red line 5.17). This suggests that experience working on the ORKG may have a positive impact on the quality of comparisons created by the same person over time.

However, it is important to keep in mind that this is only speculation based on a small sample size, and further analysis would be needed to draw any definitive conclusions.

Chapter 6

Discussion and Future Work

In this chapter, I discuss the strengths and limitations of my work in curating and analyzing the ORKG, as well as potential future directions for research. The analysis of the ORKG provides valuable insights into the quality of comparisons created by contributors and how this quality changes over time. Additionally, my work contributes to the broader efforts to develop and maintain the Open Research Knowledge Graph.

I begin by discussing the strengths of my work, highlighting the contributions and value of my analysis. I then turn to the limitations of my approach, discussing areas where my analysis may be incomplete or subject to potential biases. Finally, I conclude by outlining potential avenues for future research and development in the field the Open Research Knowledge Graph.

6.1 Strengths and Limitations

The previous chapter discussed the results of the analysis of the ORKG curation grants data, providing useful insights into the quality of comparisons created by different contributors over time. However, it is important to acknowledge that any study has its limitations and to consider them when interpreting the findings. In this section, I examine the strengths and limitations of the current study, which helps contextualize the results and inform potential future research directions.

Measuring the quality of data is a complex task that requires careful consideration of several factors, such as property completeness, instance completeness, population completeness, and semantic accuracy. In this study, I attempt to measure these factors in the ORKG dataset, which consists of almost 400 comparisons across different fields of study.

Property Completeness

The property completeness measure is an important metric that assesses the quality of a comparison by evaluating how complete it is in terms of the properties being compared. In this study, I analyze almost 400 comparisons in the ORKG curation grants to determine the property completeness score. My analysis showed that the average property completeness score across all comparisons is 80.97%, indicating that there are not many empty cells in the comparisons. This is a positive finding because it suggests that contributors are extracting similar information from different contributions, making them more comparable.

Furthermore, the high property completeness score indicates that there is a robust and comprehensive set of properties being used in the comparisons. This is important because it enhances the overall quality and reliability of the data. By having a complete set of properties, the comparisons can be more easily analyzed and interpreted, providing valuable insights for researchers and other stakeholders.

It should be noted, however, that there may be some limitations to the property completeness measure. For example, there may be certain properties that are not relevant to all contributions, leading to some empty cells. Additionally, the completeness score does not reflect the accuracy or quality of the information being provided in the comparisons. Therefore, while the high property completeness score is a positive finding, it should be interpreted alongside other measures of quality and accuracy to provide a comprehensive assessment of the data.

Instance Completeness

The instance completeness measure provides valuable insights into the quality of the data being compared. However, it is important to note that the interpretation of the score can be challenging. The approach used in this study assumes a closed set of possible values that all cells of a property can have in common. This approach works well for certain types of data, but it may not be appropriate for data where each cell has its own value. For example, in the case of measured values, each cell of a property may have a unique value, and it may not be possible to have all cells with the same value. Therefore, the calculation of instance completeness may never reach a score of 100%. It is essential to consider this limitation when interpreting the instance completeness measure.

Furthermore, my study shows that the average instance completeness score in ORKG curation grants is 40.84% across all comparisons. This suggests that there is still potential for development in terms of instance completeness. It is crucial to remember, however, that this score may not correctly reflect the completeness of the

data being compared. The present approach is based on the assumption that all contributions have the same set of instances, which is not necessarily the case. As a result, the instance completeness score should be read cautiously, and other criteria such as the complexity of the data being compared should also be taken into account.

Population Completeness

The population completeness measure is an important metric for assessing the completeness of a comparison in terms of the population being compared. However, in the case of the ORKG, there is a limitation in measuring this metric since the total population of entities is not clearly defined in the dataset.

Despite this limitation, my analysis of the population completeness measure across all comparisons in the first and second curation grants provides a valuable visualization for users and shows how data is distributed in the comparisons. This can help researchers to identify areas where further data is needed, as well as highlight areas where data is already abundant. In addition, the visualization can assist in identifying patterns and trends in the data, which can inform future research directions.

Overall, while the population completeness measure is limited in the ORKG dataset, it still provides valuable insights into the completeness of comparisons and can guide researchers in identifying areas where further data is needed to improve the accuracy and completeness of the dataset.

Semantic Accuracy

The semantic accuracy measure assesses the extent to which a comparison is accurate in terms of its semantics. This is a difficult measure to assess since it requires a deep understanding of the semantics of the data being compared. In this study, I measure the semantic accuracy at the contribution level, but surprisingly I find several issues when measuring it across all comparisons in the first and second ORKG curation grants.

Firstly, I encounter many zeros after calculating the semantic accuracy scores, indicating a lack of templates in the comparisons.

The other issue I found when measuring semantic accuracy across the comparisons is the ease with the possibility for contributors to add contributions without associating any paper information. This problem arises because contributors are allowed to create a contribution without a paper and add it to the comparison, leading to a wrong table. In one case, a contributor publishes a comparison with 12 contributions, but only 11 are shown, leading to significant inaccuracies in the data.

One possible solution to these issues is to incorporate an automated quality control mechanism that ensures that all contributions are validated against established templates and that the data is consistent across all comparisons. This would ensure that all contributions are valid and that they provide the necessary information to ensure semantic accuracy.

It's important to note that the semantic accuracy measured in this study is only at the contribution level, and does not take into account the deeper levels of the knowledge graphs in the ORKG. The depth of the graphs in the ORKG can vary significantly, and therefore, the semantic accuracy can also differ at different depths.

While the absence of a template may result in a 0 score for semantic accuracy in automatic checking, it does not necessarily imply that the comparison is semantically inaccurate. There may be cases where the comparison is indeed correct, despite receiving a low score in semantic accuracy due to the limitations of the current automatic checking approach

In addition, it's worth mentioning that measuring semantic accuracy is a challenging task due to the inherent subjectivity involved in interpreting the semantics of data. However, it is an essential measure to ensure the accuracy and reliability of data in knowledge graphs.

Overall, the semantic accuracy measure in this study provides a good starting point for improving the accuracy of data in the ORKG, but further work is needed to explore ways to measure semantic accuracy at different depths of the graphs and to address the issues identified in this study.

6.2 Conclusion

In this thesis, I present the findings of my study of the ORKG curation grants, which aim to assess the quality of the curated data using completeness and accuracy measures. My findings show that, while specific completeness metrics, such as instance and population completeness, have limits, the overall quality of the curated data is rather strong, with an average property completeness score of 80.97%. Nonetheless, there is still potential for improvement, notably in the instance and population completeness measurements, which necessitate a more thorough examination of the complications inherent in evaluating these completeness measures.

6.3 Future Work

My analysis has identified several areas for future work that could help to address the challenges and limitations associated with measuring the completeness and accuracy of knowledge graphs. These areas include:

- Developing more robust and standardized measures for evaluating the completeness and accuracy of knowledge graphs, particularly with regard to the instance and population completeness measures.
- Improving the coverage and quality of the data being curated by knowledge graphs, through collaborations with domain experts and the use of more advanced data integration techniques.
- Evaluating the effectiveness of different curation strategies and approaches for improving the completeness and accuracy of knowledge graphs, such as crowdsourcing, machine learning, and expert curation.
- Investigating the use of semantic web technologies and ontologies to improve the completeness and accuracy of knowledge graphs, through the development of more fine-grained and standardized semantic models.
- Developing tools and frameworks for visualizing and exploring the completeness and accuracy of knowledge graphs, to facilitate better understanding and interpretation of the data.

By focusing on these areas for future study, I may contribute to improving the quality and coverage of knowledge graphs, as well as developing more effective ways to evaluate their completeness and correctness. Eventually, this will aid in the support of a wide range of applications such as data integration, knowledge discovery, and decision-making, as well as drive future improvements in the field of knowledge representation and management.

Furthermore, my analysis highlights several challenges and limitations associated with measuring the completeness and accuracy of knowledge graphs. These challenges include difficulties in defining and measuring completeness and accuracy measures, as well as limitations in the quality and coverage of the data being curated. Addressing these challenges will require continued efforts to improve the quality and coverage of knowledge graphs, as well as the development of more robust and standardized measures for evaluating their completeness and accuracy.

Bibliography

[1] Dean Allemang and James Hendler. *Semantic Web for the Working Ontologist: Effective Modeling in RDFS and OWL*. Morgan Kaufmann, 2014.

[2] Sören Auer et al. "Dbpedia: A nucleus for a web of open data". In: *The Semantic Web: 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007+ ASWC 2007, Busan, Korea, November 11-15, 2007. Proceedings*. Springer. 2007, pp. 722–735.

[3] Tim Berners-Lee. "Linked data". In: *W3C Design Issues* 1.1 (2011), pp. 1–20.

[4] Jiahui Chen, Jun Zhao, and Juanzi Li. "Knowledge graph evaluation: A systematic review and future directions". In: *ACM Transactions on Knowledge Discovery from Data (TKDD)* 15.2 (2021), pp. 1–33.

[5] Gene Ontology Consortium. "The Gene Ontology (GO) database and informatics resource". In: *Nucleic acids research* 32.suppl_1 (2004), pp. D258–D261.

[6] Enrico Daga, Mathieu d'Aquin, and Alessandro Adamou. "An empirical investigation of properties of knowledge graph embeddings for link prediction". In: *The Semantic Web–ISWC 2019*. Springer. 2019, pp. 99–117.

[7] Nan Du et al. "Towards instance-level entity alignment between knowledge graphs". In: *Knowledge-Based Systems* 171 (2019), pp. 78–90.

[8] Jérôme Euzenat and Pavel Shvaiko. "The ontology of knowledge graphs". In: *Semantic Web* 7.3 (2016), pp. 205–215.

[9] Jérôme Euzenat et al. "OAEI 2020 campaign". In: *Proceedings of the 15th International Workshop on Ontology Matching co-located with the 19th International Semantic Web Conference (ISWC 2020)*. 2020, pp. 101–132.

[10] Catherine Faron-Zucker et al. "KG-BENCH: A Benchmarking Framework for Knowledge Graph Systems". In: *The Semantic Web - ISWC 2021*. Springer. 2021, pp. 351–366.

[11] Andrés Garcıa-Silva, Oscar Corcho, and Raúl Garcıa-Castro. "Semantic accuracy assessment in ontologies". In: *Expert Systems with Applications* 39.14 (2012), pp. 11767–11779.

[12] Charles R Harris et al. "Array programming with NumPy". In: *Nature* 585.7825 (2020), pp. 357–362.

[13] Wajdi Hassan, Tommaso Di Noia, and Eugenio Di Sciascio. "Towards a knowledge graph-based framework for product comparison". In: *International Conference on Advanced Information Systems Engineering*. Springer. 2019, pp. 36–50.

[14] K A Heckert and D M Louzao. "Challenges in Defining the Relevant Population for a Clinical Question". In: *Journal of hospital librarianship* 17.1 (2017), pp. 1–7.

[15] Sebastian Hellmann et al. "Towards Knowledge Graph Quality Benchmarks". In: *Proceedings of the 18th International Semantic Web Conference (ISWC 2019): Posters and Demonstrations*. 2019.

[16] Laura Hollink, Krisztian Balog, and Giuseppe Rizzo. "Assessing the quality of knowledge graphs: A systematic review of frameworks, tools, methodologies, and metrics". In: *Proceedings of the 30th ACM Conference on Hypertext and Social Media*. ACM. 2021, pp. 125–136.

[17] Naimdjon Huma, Jens Lehmann, and Sören Auer. "Interactive quality assessment of knowledge graph completeness and consistency". In: *The Semantic Web–ISWC 2020*. Springer. 2020, pp. 344–361.

[18] John D Hunter. "Matplotlib: A 2D graphics environment". In: *Computing in science & engineering* 9.3 (2007), pp. 90–95.

[19] Atif Hussein, Christoph Lange, and S"oren Auer. "Knowledge Graph Maturity Model (KGMM): Guidelines for Growing and Evaluating Scholarly Knowledge Graphs". In: *European Semantic Web Conference*. Springer. 2021, pp. 63–79.

[20] Hassan Hussein et al. "KGMM-A Maturity Model for Scholarly Knowledge Graphs Based on Intertwined Human-Machine Collaboration". In: *From Born-Physical to Born-Virtual: Augmenting Intelligence in Digital Libraries: 24th International Conference on Asian Digital Libraries, ICADL 2022, Hanoi, Vietnam, November 30–December 2, 2022, Proceedings*. Springer. 2022, pp. 253–269.

[21] Leo Jansen, Lourens van der Meij, and Mark Graus. "Critical Success Factors for Implementing Knowledge Graphs: A Systematic Literature Review". In: *IEEE Access* 9 (2021), pp. 2533–2553.

[22] David R Karger, Sewoong Oh, and Devavrat Shah. "Towards Robust Metrics and Benchmarks for Entity Resolution". In: *Proceedings of the 2011 International Conference on Management of Data*. ACM. 2011, pp. 145–156.

[23] Abhijit Kulkarni, Mayank Singh, and Praveen K Paritosh. "CrowdER: Crowdsourcing-based Entity Resolution in Knowledge Graphs". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2019, pp. 6109–6119.

[24] Sheng Liu, Jingbo Zhang, and Yu Zheng. "Populating Knowledge Graphs: A Survey of Techniques and Resources". In: *ACM Transactions on Knowledge Discovery from Data (TKDD)* 14.4 (2020), pp. 1–41.

[25] Xiaoxiao Ma et al. "BioBench: a benchmark suite for evaluating biomedical knowledge graph construction and quality". In: *BMC Bioinformatics* 22.1 (2021), pp. 1–14.

[26] Wes McKinney. "Data structures for statistical computing in python". In: *Proceedings of the 9th Python in Science Conference* (2010), pp. 51–56.

[27] Dat Quoc Nguyen, Dai Quoc Nguyen, and Thanh Vu Nguyen. "Knowledge graph embedding via dynamic mapping matrix". In: *Proceedings of the 26th International Conference on World Wide Web*. 2017, pp. 687–696.

[28] T Nguyen et al. "Expanding and refining benchmarking sets for protein functional annotation". In: *Proceedings of the Fifth BioCreative Challenge Evaluation Workshop*. 2015, pp. 28–33.

[29] Harshvardhan J Pandit et al. "Luzzu—An Extensible Linked Data Quality Assessment Framework". In: *Proceedings of the 19th International Semantic Web Conference (ISWC 2020) (Posters & Demonstrations Track) (ISWC 2020 Posters & Demonstrations Track) (ISWC 2020)*. International Semantic Web Conference. 2020, pp. 1–4.

[30] Heiko Paulheim. "Knowledge graph refinement: A survey of approaches and evaluation methods". In: *Semantic Web*. 2017, pp. 489–508. DOI: 10.3233/SW-160218.

[31] Heiko Paulheim. "Knowledge graphs". In: *Journal of Data and Information Quality (JDIQ)* 8.4 (2017), pp. 1–14.

[32] Harald Sack et al. "The Knowledge Graph Quality Assessment Challenge". In: *Proceedings of the 18th International Conference on Knowledge Capture*. ACM. 2019, pp. 1–8.

[33] Max Schmachtenberg, Klemens Böhm, and Thomas Kirste. "Assessment of quality flaws in geospatial ontologies: The case of GeoNames". In: *International Conference on Knowledge Engineering and Knowledge Management*. Springer. 2016, pp. 185–200.

[34] Kuldeep Singh, Prabhakar Rajagopal, and Jens Lehmann. "KG-Benchmarker: Towards a Comprehensive Benchmarking Framework for Knowledge Graph Systems". In: *Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM '21)*. ACM. 2021, p. 10.

[35] Claus Stadler, Duy Nguyen, and Johanna Völker. "Completeness Tradeoffs in Ontology-Driven Knowledge Graph Population". In: *arXiv preprint arXiv:2010.15271* (2020).

[36] Xiaoyu Sun et al. "Crowd-verification: A crowdsourcing approach to verify knowledge graph facts". In: *Journal of Web Semantics* 68 (2021), p. 100595.

[37] Andreas Trameşberger et al. "Towards Instance-based Evaluation of Knowledge Graph Population". In: *arXiv preprint arXiv:2012.15794* (2020).

[38] Marieke Van Erp et al. "Assessing the completeness of knowledge graphs for answering biomedical questions". In: *Proceedings of the 2nd Joint Workshop on Deep Learning for Biomedical Natural Language Processing and Biomedical Knowledge Graphs*. 2020, pp. 15–25.

[39] Xiaowei Wang et al. "Detecting Erroneous Relations in Knowledge Graphs with Machine Learning". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2020, pp. 4028–4038.

[40] Xinyue Wang et al. "Quality Assessment and Error Correction of Knowledge Graphs: A Review". In: *IEEE Access* 9 (2021), pp. 21644–21660.

[41] Martin White. *Knowledge Graphs: An Introduction*. Morgan Claypool Publishers, 2019.

[42] Mark D Wilkinson et al. "The FAIR Guiding Principles for scientific data management and stewardship". In: *Scientific Data* 3.1 (2016), pp. 1–9.

[43] Zheng Xie et al. "Instance level completeness evaluation for knowledge graph". In: *2018 IEEE International Conference on Big Data (Big Data)*. IEEE. 2018, pp. 3035–3042.

[44] Amrapali Zaveri et al. “Quality assessment for linked data: A survey”. In: *Semantic Web* 7.1 (2016), pp. 63–93.

[45] Fuzheng Zhang et al. “A comprehensive survey of knowledge graphs: Representation, acquisition and applications”. In: *IEEE Transactions on Knowledge and Data Engineering* 30.2 (2018), pp. 312–333.

[46] Fuzheng Zhang et al. “Enhancing Knowledge Graph Completion with Jointly Learned Linguistic Patterns and Triples”. In: *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM. 2019, pp. 385–394.